Dengue Prediction Model

A COMPARATIVE STUDY OF MACHINE LEARNING TECHNIQUES

Tarun Swarup

2988527

Submitted in partial fulfilment for the degree of

Master of Science in Big Data Management & Analytics

Griffith College Dublin
September, 2019

Under the supervision of Supervisor's Name
Abubakr Sidig

Disclaimer

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the Degree of Master of Science in Applied Digital Media at Griffith College Dublin, is entirely my own work and has not been submitted for assessment for an academic purpose at this or any other academic institution other than in partial fulfilment of the requirements of that stated above.

Signed: _____ Date: _____

Abstract

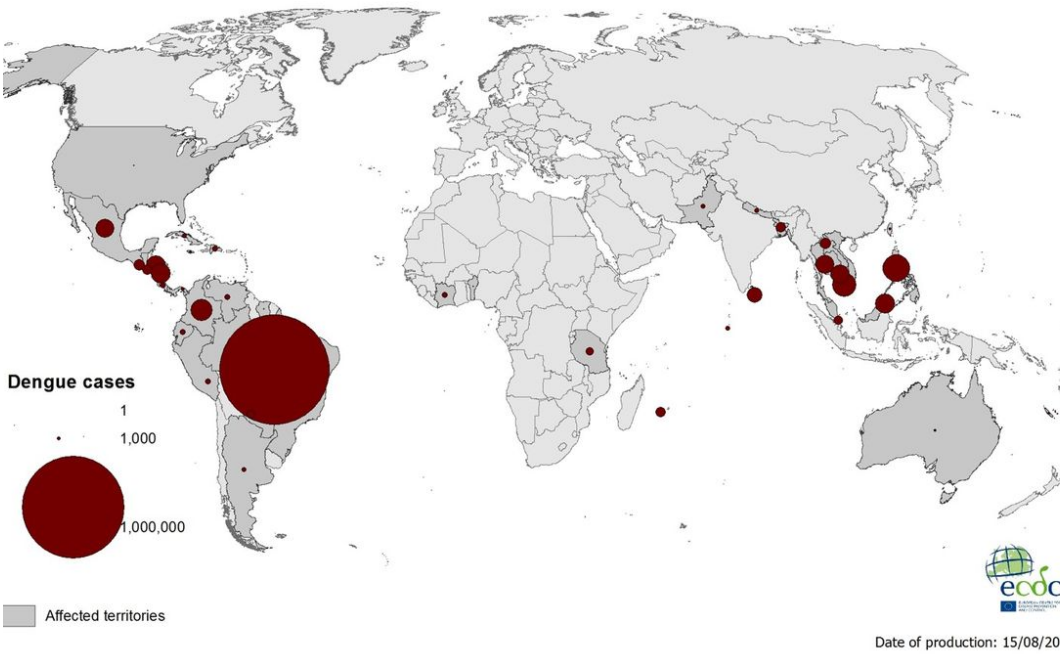
Machine learning is nowadays the trend in the field of computer science. The power of machine learning combined with various sophisticated technologies developed such as Internet of Things , data analytics , Big Data , Data science , block chain is unfathomable. The knowledge one gains from these domains is invaluable. Healthcare is also influenced and benefitted by machine learning algorithms. To see the future can be really challenging , but has become possible with the aid of such technologies. This paper deals with Dengue and thrives to develop a model that would predict the occurrence of dengue in a patient , provided the necessary symptoms are fed in. Random Forest Classifier , Support Vector Machine , Logistic Regression are the machine learning algorithm we will be studying and implementing on a Dataset containing the characteristic symptoms of Dengue. The aim is also to apply this predicted set in an IoT platform so that the user involved can know the presence of the disease based on the input values given .

Chapter 1

Introduction

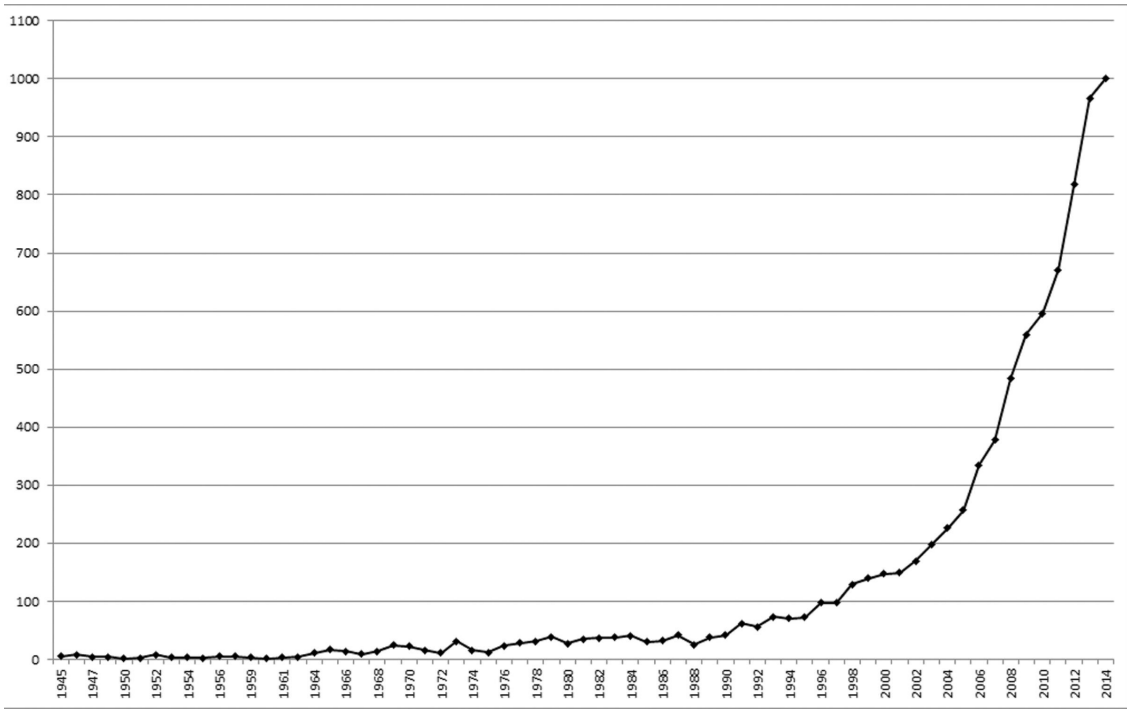
1.1 DENGUE : AN OVERVIEW

Dengue is a pandemic disease prevalent in various regions of the world. The disease is found to be affecting people throughout the world but it predominantly develops in tropical and subtropical areas. It occurs extensively in parts of Africa , South America , Carribean and SouthEast Asia [1]. It can be found in urban , peri-urban areas and suburbs as well , where the population is much dense[2].



DENGUE : A WORLD WIDE MAP

The incidence of this disease has grown up to a large extent in the past two decades [3] .It is a mosquito-borne diseases caused by four different types of virus. People from over 120 countries have found to be affected by the disease. It is life-threatening and has not ended the global outspread throughout the years. Studies say that dengue can be present anywhere in tropical regions with the risk level dependent on the degree of rainfall , climate and urbanisation.The Dengue fever is recognised as both a global endemic and epidemic disease. The symptoms are somewhat similar to diseases like typhoid and malaria but still there is no actual cure to the disease [4]



INCREASE OF DENGUE IN GLOBAL POPULATION : A TIMELINE

The aim of this Masters Thesis is to investigate and implement algorithms that could possibly predict the occurrence of dengue in humans. The major factors of the disease are to be taken into account to predict the infection. Data mining techniques and machine learning algorithms can be used for the prediction of future happenings. There exist several algorithms in machine learning to develop a disease prediction model such as Linear and Logistic Regression , Decision Tree and Support Vector Machine (SVM) . The plan is to predict the occurrence of the disease in humans by a machine learning prediction model trained using Support Vector Machine as well as Logistic Regression algorithms in order to achieve the result with a good accuracy rate. A medical dataset with state wise records of patients from India is to be fed into the model , depending on which the result can be positive or negative. Attributes include general information about the patient

, physical factors like muscle pain , fever , fatigue , eye pain , temperature , blood platelet count etc. A detailed description about the dataset used is given in Chapter 3. We aim to build a model that achieves an accuracy rate above 85%.

1.2 GOALS

The proposed model would be highly helpful and beneficial in the medical field to diagnose the disease in advance and help doctors prevent it from occurring. Further , by integrating IoT (Internet Of Things) to the model , we can send SMS alerts to the the family members or the doctor of a potential dengue patient (simulated). The predicted result , which is the output from the previous phase is uploaded to an IoT platform so that , when real time data is sent , the presence of the disease can be identified as well as alerted to any given emergency contact. The text message contains information about the patient such as GPS data (Latitude , Longitude) and whether dengue is positive or negative in them. The design and implementation of the model is discussed in detail throughout Chapter 4. Third party IoT and bulk sms integration tools snd platforms were used in order to achieve the goals. The figure below shows the number of cases and deaths faced by Indian population from 2012 to 2019.

DENGUE IN INDIA

1.3 MOTIVATION

This project idea was motivated keeping in mind the figures and facts about dengue and the rapid growth of the viral disease around the world , mainly South America , Carribean and South East Asia , out of which India is a well known host of the virus. Firstly , found in Western Part of Bengal (Calcutta) but the disease has been identified in most of the India states except the regions of Kashmir and Himalayas. The fever is a major concern of public health due to its sudden occurrence and the rate of disease in the country's population [5]. Moreover , it is an infection with high mortality rate in India and with no proper treatment or cure.

1.4 Method of Approach

The first step was to explore various datasets available in the internet related to the project. It was a bit challenging to choose the most appropriate one in the start. The dataset selected for our project is a health and medical dataset consisting of medical records of patients from a number of states . The list of attributes include factors like fatigue , muscle pain , skin rash , eye pain etc of different patients . These were needed in order to test and train our model for prediction of the disease.

The project implementation was broken down into four phases like training the model , integrating Internet Of Things to the model to detect the disease (say data was considered from simulating a patient) , prediction and sending alerts based on the results (text message). The model was trained and validated using Python Programming Language in Jupyter notebook . The IoT platform used was Ubidots and the third party SMS integration platform called TextLocal was included as well.

Various resources from the internet were studied and analysed in order to build the project flow , implement methods and functionalities and to arrive at an acceptable result. A wide range of References including Research papers , science journals , government and university websites related to dengue fever , machine learning , Internet Of Things were studied . The summary of all the references used are discussed in Chapter 2 . (Background work) The methodologies given under Chapter 3 include how then project was approached and accomplished , technologies , platforms and applications used in our project. The basic specifications required to execute our

project and the technical aspects dealt with during the course of the project are outlined in the Chapter 4. The design and implementation part of the project are discussed under Chapter 5 . The project implementation consists of the work breakdown , the range of techniques and functionalities applied ,and the operations performed to bring about our project are talked through in this Chapter . The results and sequential outcomes through the testing and evaluation of our project are discussed in chapter 6. This also educates the reader of the challenges and possible difficulties faced in the whole process right from choosing our source dataset to predicting the result. The conclusions drawn from the project , advancements that could be made to the current design , future work and interests are shared in the final chapter.

Chapter 2

Background

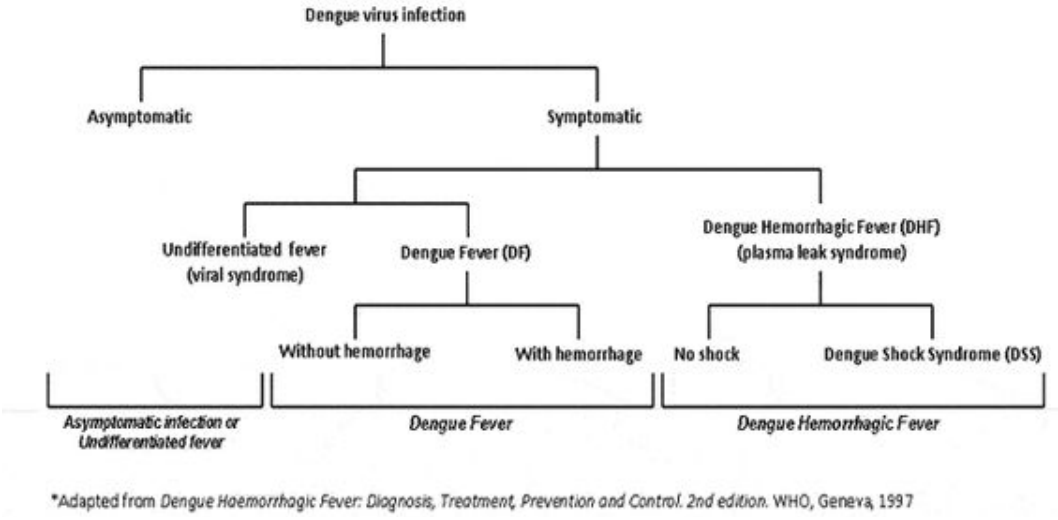
2.1 REVIEW OF LITERATURE

2.1.1 DENGUE : HISTORY

Initial occurrences of the Dengue fever dates back to the Medieval Period. A Chinese encyclopaedia consisting of symptoms and treatments of diseases was found to have the first and foremost record of an infection caused by water-borne flying insects . The book was formally written by Chinese dynasties in the Middle Ages. (265 to 992 AD) They termed the disease as water poison. Widespread occurrence of dengue-like infections were in the regions of Asia , Africa and American continents. (2.1) The condition of dengue fever was widely distributed in tropical and sub-tropical urban areas. Dengue is caused by either of the four virus serotypes (DEN-1 , DEN-2 , DEN-3 and DEN-4) belonging to genus Flavivirus. A Parasitic species of mosquito , Aedes aegypti and Aedes albopictus are the two vectors of the virus , that leads to dengue . In 1995 , Dengue was recognised as the most important mosquito borne viral illness affecting , both morbidity and mortality of the human race.(2.2) The disease was widely known by certain pseudonyms namely break-bone fever , dandy fever , denguero , bouquet fever , polka fever or 7-day fever. Records say that about two million people from the Southern states of America were infected with the dengue fever. (2.3)

2.1.2 DENGUE : SYMPTOMS

Indications of dengue fever outbreak in the human body , a week after the bite of the mosquito bearing the virus , characterised by fever , fatigue , eye pain , muscle pain , joint pain and bone pain. These symptoms lead to the further weeks of the fever where the affected person could have skin rash similar to measles , damage to lymph nodes and mild haemorrhage. (2.4) the basic dengue fever is asymptomatic at times.

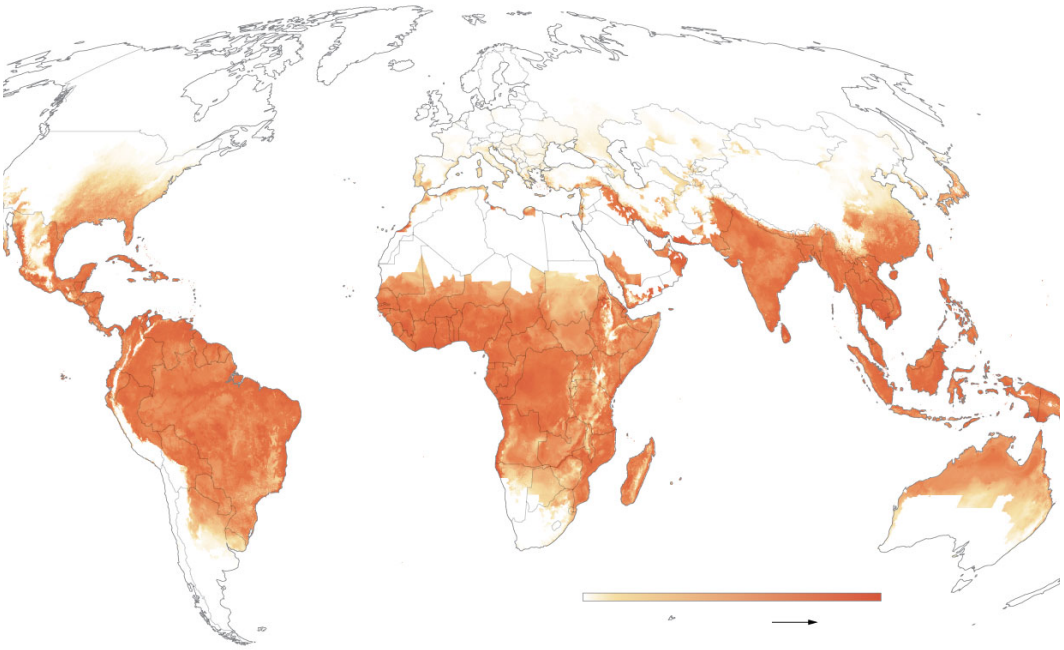


DENGUE : COMPLEXITY

2.1.3 DHF AND DSS

Dengue Hemorrhagic Fever (DHF) and Dengue Shock Syndrome (DSS) are two life-threatening diseases caused by the dengue virus. Approximately , five percent of the population are noted to have these fatal diseases from 50 million dengue cases every year. (2.5) They are considered to been of the major reasons of hospitalisation and mortality. The following are some discussions on complications and clinical manifestations that could be identified in patients. Dengue Hemorrhagic fever is seen with 4 main symptoms : severe fever , haemorrhage (damage of blood vessels) , enlargement of the liver , leakage of plasma ; eventually leading to the failure of circulatory system. Children affected by the disease face anorexia , nausea , abdominal pain and muscle pain. When plasma loss is over critical , it develops shock and death if left untreated. This condition is known as Dengue Shock Syndrome (DSS) . The severity can be weakened by early replacements of plasma in the body. It comes along with pulse pressure narrowing down rapidly (hypotension). Patients are conscious even in the critical stage of shock but with no appropriate therapy , they face death. (2.7)

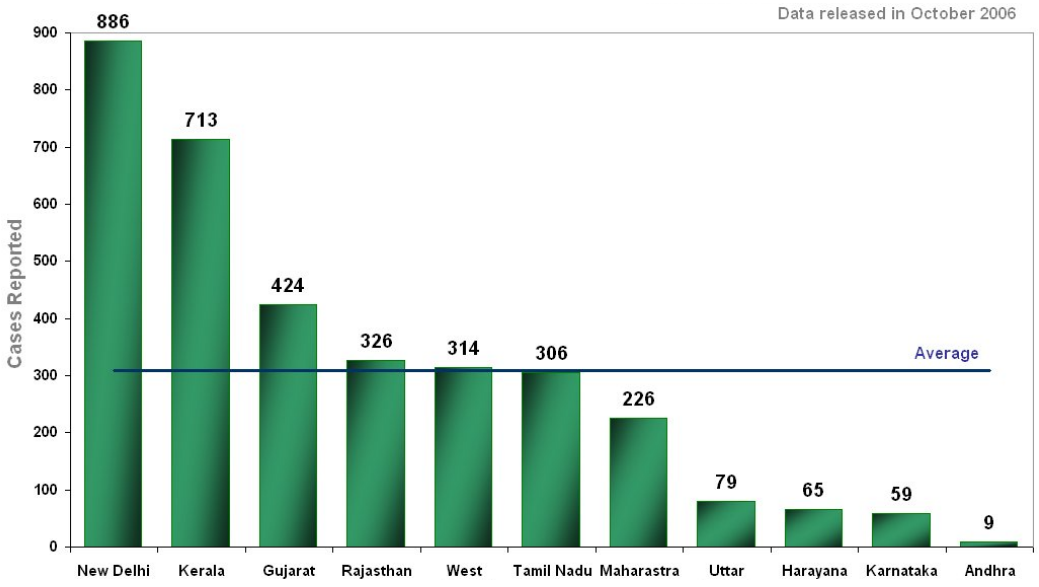
The global mortality and morbidity rates are greatly influenced by the amount of dengue cases that exist in various geographical locations. Health experts have manipulated numerous approaches to map dengue risk and moderated possible endemics , considering the history of the disease and its occurrences worldwide. The World Health Organization had estimated apparent figures of the infections globally . In 2010 , it was roughly calculated that Asia contributes to 70% of the world's dengue burden , America 14 % and Africa 16%. The whole of India bore significantly higher burden than most of the Asian countries (34 %) . Dengue infections were predicted by deriving relationships between the risk of the disease and its incidence by the cohort studies. (2.8) A better understanding of the disease and its epidemiology (branch of medicine dealing with incidence , distribution and possible control of the disease) will require deeper analysis and prospective studies. The origin of the disease and the virus is still unknown though studies suggest that it is Africa as the vector mosquito species , Aedes aegypti who bear the virus are believed to exist there in huge amounts. (2.9) Sylvatic research about the disease shows us that all the four virus serotypes exist in Asia. (2.10)



FORECAST 2080

2.1.4 DENGUE IN INDIA

The fever prevails in mostly all tropical and sub-tropical areas. All four types of virus causing the disease circulate in India which is the reason for dengue epidemics there . Previous medical records prove that India has an intense history with dengue. WHO published a map to show the distributions of epidemic activity of the disease in 2006 where India was marked severe. Dengue in India has always been resurgent because of factors like uncontrolled population , urbanisation and insufficient infrastructure of public health. It is somewhat natural that the species propagate in such a high densely populated nation. After the Second World War , the virus gained its increase in transmission , due to sudden urbanisation and outgrowth of population in SouthEast Asia. An illness similar to dengue was reported in the Southern part of India in 1780. (2.11)



DENGUE CASES REPORTED IN INDIA - 2006

Delhi , the national capital of India situated in its Northern part had numerous outbreaks of the virus in the late 90's ; the largest being in 1996. 8900 cases of the disease were identified with a mortality rate of 4.2%. Serological tests of blood samples were carried out in the All India Institute of Medical Sciences (AIIMS) collected from patients https://apps.who.int/iris/contact in and around the region. (2.12) This was considered as the largest reported endemic from the Indian region. Various regions are potential to add to the country's epidemic. Several proposals were made and experiments were performed in patients to give explanation to the pathogenesis (the manner of development of the disease) of the DHF fever to forecast the occurrence of the disease and minimise the country's burden and human agony.

	2014		2015		2016		2017	
	Cases	Deaths	Cases	Deaths	Cases	Deaths	Cases	Deaths
Kerala	2575	11	4075	25	7439	13	18,727	35
Karnataka	3358	2	5077	9	6083	8	13,016	5
Tamil Nadu	2804	3	4535	12	2531	5	11,552	18*
West Bengal	3934	4	8516	14	22,865	45	5389	13
Delhi	995	3	15,867	60	4431	10	4545	1

DENGUE IN SOUTH INDIA

There are very few laboratories in the country back then which actually had virology facilities to diagnose diseases. Later , the disease became pandemic in the whole of India with the virus group being active and ubiquitous. The WHO state that dengue is still not severely endemic in the European continent. The European Union (EU) observes that even the Incidents previously reported were due to tourists travelling from overseas and other endemic territories. (European Centre for Disease Prevention and Control 2.14) Tourists play a significant role in the disease's epidemiology around the world as they transport serotypes to places with mosquitoes and further spread the infection.

The Aedes mosquito species are believed to circulate between latitudes 35N and 35S and they are comparatively more prevalent at the elevation of 1000 metres or below. (2.13)

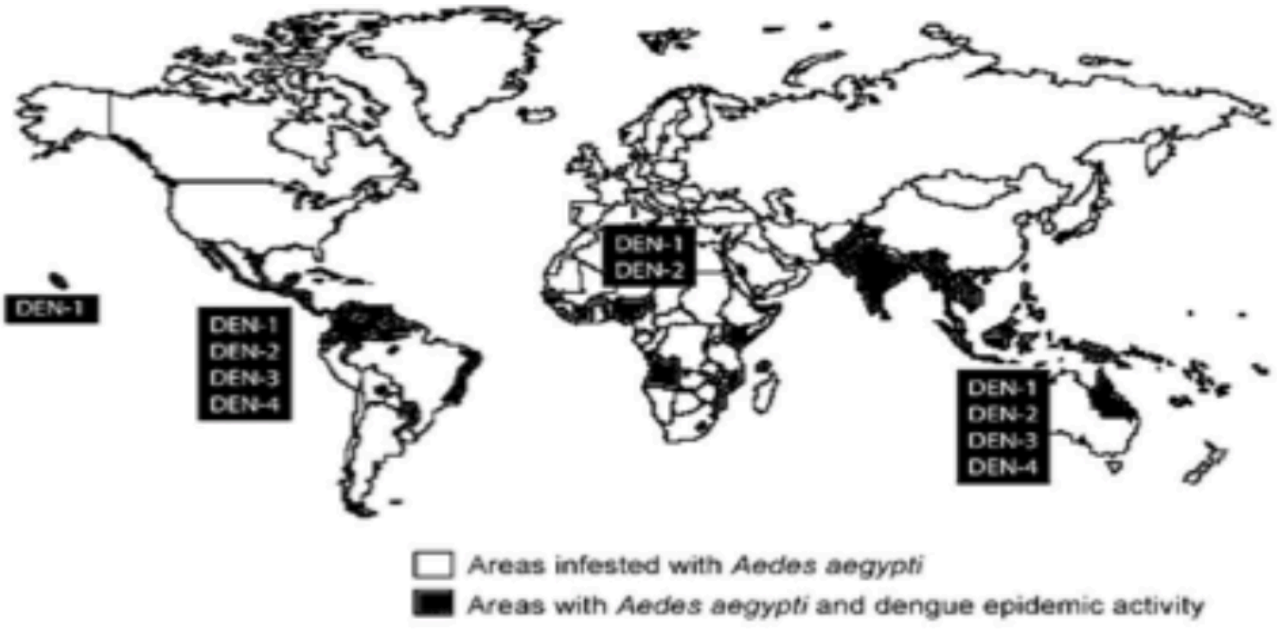
The reason behind the transmission of the arbovirus (a group of viruses transmitted by mosquitoes , ticks or other arthropods) includes several factors such as environment , climate , interactions between host and pathogens (virus causing the disease) and immunity of the population .

Prevention

Minimising the transmission of dengue serotypes is totally conditional on controlling the vector species and reducing the contact between the host and pathogens. The WHO encourages a strategic approach towards dengue vector control known as Integrated Vector Management (IVM) , which was stated as " a rational decision-making process for the optimal use of resources for vector control ". The prevention of the disease requires deeper understanding of the vector's relationship with the environment and its natural habitat. (2.16)

Mobilisation of people should be encouraged along with administrative , legal and regulatory advocacy. (13) Sanitary legislation and modification of housing and water storage systems should be imposed on the responsible organisations. (2.17) In India , several legislative measures and dengue programmes were regulated in order to detect and control the breeding of mosquitoes , which led to a significant emphasis on dengue prevention. (2.18)

The DHF/DF has de facto become hyper-endemic in most of the tropical and sub-tropical urban regions. The South- Eastern part of Asia accounts for more than half of the world's risk due to dengue and its complications. DengueNet is a data management scheme to analyse and survey the



epidemiological data and records in order to further suppress and diminish the spread of Dengue regionally as well as globally. The prime goals were to nourish disease surveillance . The consensus was to identify and improve virological laboratories , learn more from related vector-borne diseases like malaria and integrate health workers and inspectors with respect to the Integrated Disease Surveillance Programme. (IDSP) (2.19) The major States like Delhi , Tamil Nadu , Karnataka , Maharashtra and Uttaranchal participated in the meetings and collaborations conducted by WHO . Due to previous instances of dengue outbreaks , public laboratory services were operated as tie-ups with the organisation.

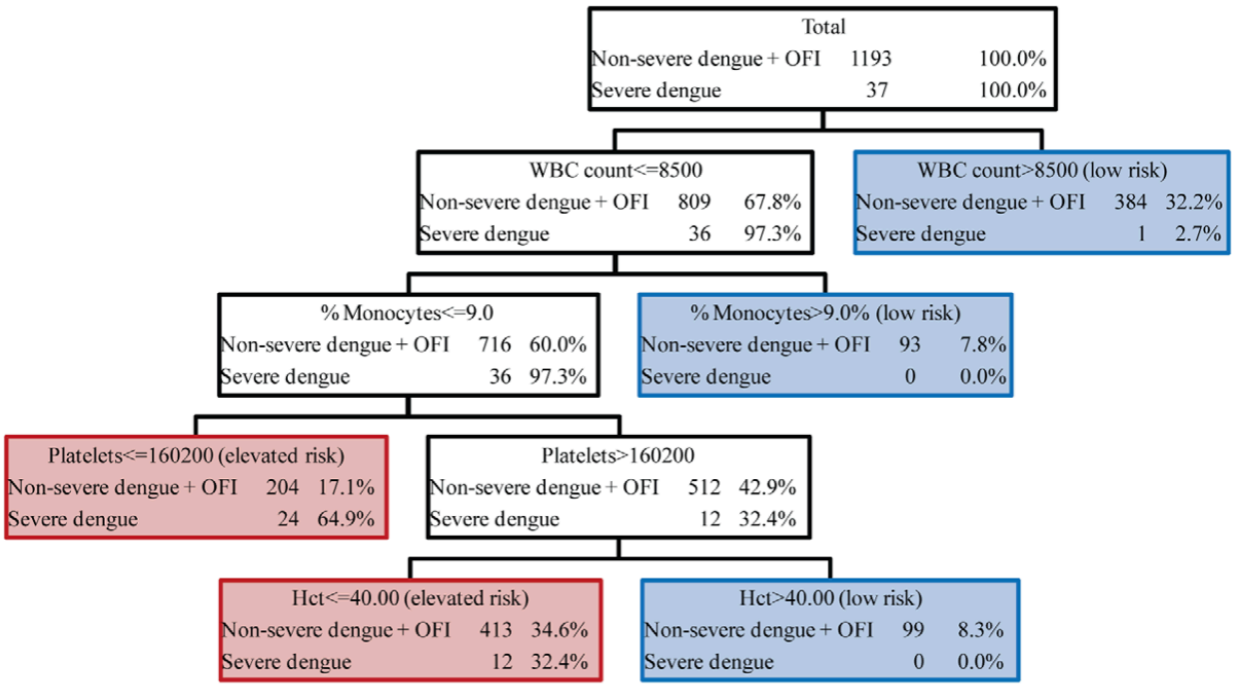
If a person is infected with either of the four dengue serotypes , they become less immune develop the risk of being attacked by the other three viruses , which might result in Dengue Hemorrhagic Fever (DHF). (2.20) Several prognostic tests and analysis are carried out to detect the severity of the virus in early stages of illness. Medical researchers say that enhancement of antibodies and their contribution in the Dengue fever complexities can be aroused by Tetravalent dengue vaccines , which tranquillize the response to all four serotypes. Drugs are also designed with compounds that act against the dengue virus and block their replication. There are numerous vaccines developed and experimented like TV003 , TDV (DENVax) despite the limitations in R&D and investment in vaccines , but there is still no actual cure for the disease. Greater effort should be taken to vaccinate children to develop immunity of the virus in the later stages. (2.21) An effective and advanced dengue vaccine should be studied and discovered by WHO to reduce the mortality and morbidity of the disease by 2020. Further advancements in approaches like Mathematical modelling , prognostic assays , therapies , insecticides and vaccines , dynamics and kinetics of the serotypes and better environmental modifications would collectively help in reducing the severity , burden and diminish the dengue problem. (2.22)

2.2 RELATED WORK

This section acts as a collected source of information for the reader to understand and familiarise the previous work related to the field of study. The spread of dengue fever across Singapore was studied by a group of students from Nanyang Technological University. They developed a mathematical simulation model using differential equations , considering the contemporary epidemiology and disease trend in the population. (2.23) Data was collected from people affected by dengue virus in the 1996 outbreak of Taiwan . Correlation between various clusters were analysed using Random Effect Model with a Generalised Estimation Equation to correlate the data after examination. The hypothesis showed that dengue had significant association

with the presence of water bodies , uninhabited houses , trash or other residents . Basically , environmental and household factors which are potential breeding grounds for mosquitoes should be monitored for better control. (2.24) Research papers show that epidemiological data were visualised and analysed against dengue records using different mapping and spatial modelling techniques to be incorporated in dengue programmes. (2.25)

Biomedical students from Malaysia proposed a prediction system model to forecast the time of defervescence of febrility in dengue patients without involving any sort of medical or clinical instruments in the human body . They trained and implemented the model using machine learning techniques like Artificial Neural Networks (ANN) , Multilayer Feed Forward Neural Network (MFNN) , back propagation and Multi Layer Perceptron . The model delivered an accuracy rate of 81.3% in successive iterations. Then , after applying appropriate pruning techniques , the prediction model improved to a higher accuracy rate of about 90% to predict the fever. (2.26) People affected by dengue develop characteristic symptoms and c complications only in the febrile stage of the disease. Hence , dengue usually stays asymptomatic in the starting stages even without the victim knowing about the infection. Currently , the disease can't be treated properly as there is no licensed vaccine that works against the virus serotypes. Patients who are prone to the disease are generally hospitalised to give adequate support and treatment. In addition , the essential diagnosis to characterise the severity or acuteness of the infection is not too clear. This paper published in the Neglected Tropical Diseases journal by researchers of Medical Sciences from Thailand and Massachusetts established a prediction tool to identify the severeness in the person and enhance the



hospitalisation facilities in the respective origins. The research was carried out on young children from Thailand who were infected by the illness. Data was collected periodically to learn about their symptoms and identify consecutive changes . CART (Classification and Regression Trees) Analysis was done to establish inter-variable correlation between the given attributes. Patients were then classified based on risk and sensitivity. Further , Classification trees for patients with elevated risk were generated using data read from blood samples. (2.27)

This prediction model proposed by a group of university students was developed using original data given by the National Environment Agency (NEA) in Singapore. Major goal was to apply suitable neural network techniques on dengue data recorded from actual cases to check if the results were of acceptable standards. They trained the model based on spatial parameters like temperature , humidity and rainfall in that locality. The prediction was possible dengue outcomes. The estimates were not good enough as data was only taken for a period of five weeks. (2.28) Regression algorithms are preferred over other algorithms when disease outbreaks are predicted. The experiments were performed to compare the results of two models , fuse them and understand the output.

A neural network model and a non linear regression model were built and tested to predict possible outbreaks of the disease , considering time series data from dengue cases , where the former model produced a more relevant result relatively.(2.29) Support Vector Machine (SVM) , K-H model , neural networks , time series analysis and many other techniques are used to understand the incidence of dengue globally. DHF Data of various epidemics was accumulated from reliable medical sources . Multivariate Poisson Regression gave the ranked association of different variables with respect to the number of records. The model was chosen based on Akaike’s information criterion (AIC) , Bayesian information criterion (BIC), and the mean absolute percentage error (MAPE) . Thus , the system was deemed effective with the purpose of forecasting and preventing further occurrences of infection for the people. (2.30)

Search queries from Baidu , meteorological data and weekly data samples of dengue instances in Southern China were collected. Techniques like Support Vector Regression (SVR) , Step-down linear regression , gradient boosted regression tree algorithm (GBM) , negative binomial regression , Least absolute shrinkage and selection operator (LASSO) and Generalised additive model (GAM) were used to train and validate the model to predict dengue. The SVR

model surpassed all the other variants in terms of better accuracy and minimal error rates. This was recognised as a successful effort for true dengue prediction using existing techniques. (2.31) Google designed a reporting system (query based) to detect trends in infections and diseases. These when evaluated against linear models , we derive at a good association , correlation and high accuracy with factors influencing the disease.

Google Dengue Trends (GDT) supervise search patterns over the internet and retrieve / cluster queries periodically to guess the rate of incidence. The same were examined using official data from Venezuela provided by Ministry of Health (MH) . Data was available of both peak and non-epidemic periods. Simple Linear regression was applied to calculate the overall Pearson's coefficient . High correlation was noted between Google's epidemic data and actual medical data which helped to understand the disease's trend. (2.32) Rainfall data , power supply data and dengue case data from regional / central surveillance units (CSU) and Integrated Disease Surveillance Project (IDSP) were gathered from tropical states of India from the South, Tamil Nadu and Puducherry . Analysis of Variance (ANOVA) to identify the amount of difference among the independent variables. The estimates were that spatial variations play a crucial role in transmission of the disease. Dengue has become a public health concern over the Southern states due to rapid urbanisation and ineffective preventive measures. (2.33)

Dengue Incidence maps and DALY were generated for Odisha , a densely populated region in India. The disability adjusted life year (DALY) is the parameter given by WHO to calculate the quantity of burden of the disease from mortality and morbidity for the region. Sex ratio , population density , climatic and physical factors were highly related to dengue burden. (2.34) Classification algorithms contribute significantly in the medical field to study patient's medical records and predict the existence or outcome of the disease. Students from Delhi where a major outbreak of dengue occurred developed a model to diagnose the disease in its early stages. They compared various decision tree algorithms to decide the most efficient one. J48 classifier , Random tree , REP tree , SOM and Naïve Bayes classifier were all tested with data from different hospitals. The REP tree proved to be the most accurate classifier with an efficiency of 82.7% for early detection and to provide diagnosis. (2.35) This paper extends the research to predict the propagation of Dengue Hemorrhagic Fever in Indonesian region using clustering (k-means algorithm) and classifier (Support Vector Machine) using the medical history of dengue. SVM works great with multi dimensional data. The clustering showed better accuracy of above 80%. (2.36) Prediction of any disease / infection in real-time is still nascent , but forecasts produced by researchers using different

algorithms in small scale / large scale can be of great value for global dengue risk control. Real time data from Thailand was securely transferred to the researchers in the US. Spearman Correlation Coefficient and Mean Absolute Error were calculated against real time predictions and true occurrences of the disease. The model was based on auto-regressive integrated moving average (ARIMA) . The authors prove evidence to the challenges faced when experimenting prediction of infections using real data. Hence , state that there should be deeper research done based on forecasting of diseases and along with the epidemiological knowledge of the disease , better prevention and control measures could be taken.(2.37)

A personal mobile system to alert the presence of dengue disease was proposed. The physical Data such as temperature and pressure could be drawn from Aedes kit . The data is stored in Microsoft Azure Database . The readings are observed periodically and when critical , the data is stored in cloud as well as the doctor is alerted. From an overall perspective , the model seems cost effective and could be really helpful to prevent outbreaks when actioned properly. (2.38) A Mobile application was proposed that would give the real time information about the outbreak using GPS receiver and dengue risk index around the location. The risk index was calculated by fuzzy logic reasoning. The user is categorised with a risk index of high , intermediate or low with respect to data uploaded in the database server. The working model of this proposal would be very beneficial and it also detects the major dengue hotspots eventually. Recent researches describe the application of Internet Of Things in healthcare monitoring and surveillance. (2.39) This research paper presented a model to predict dengue based on real time data and machine learning. Physical Data of the user such as heart rate , blood pressure , temperature could be collected from respective IoT sensors integrated into the living environment of the user. This information is accumulated as a dataset . Naïve Bayes classifier was said to be used to calculate posterior probability (statistical probability that a hypothesis is true calculated in with all relevant information) and the state of the disease is reported.(2.40)

IoT is a state-of-the-art technology that could possibly serve as an essential platform to diagnose and prevent vector-borne diseases like dengue. The author suggests an IoT based healthcare framework and discusses about the concept of engineering an environment to network the patient and doctor . IoT sensors , servers , cloud platforms and mobile applications can all be integrated to improve healthcare. (2.41) The collaboration of computer science with epidemiology can improve healthcare and disease prediction to a great extent. Statistical models , machine learning techniques and data representation are the fields to investigate further. The data to be used

for disease prediction should be processed and transformed into a structured format in order to arrive at a reliable result. Dynamic pandemic data about dengue should also be updated to current technologies to ensure that data is collected on a continual basis. Major challenges addressed in this field were dynamic forecasting of the disease , handling uncertain situations , big data , storage and security of patient's health data and processing to feed into disease prediction models.(2.42)

Chapter 3

Methodology

This chapter introduces the reader to different implementations used in the project . methods and techniques used to carry out the implementations , technologies and approaches followed to tackle the challenges faced throughout. The main idea of this project was to predict the presence of infection in a person based on the physical factors and symptoms identified . The factors considered were body temperature , muscle pain , presence of eye pain , fatigue , skin rash , fever and blood platelet count. A model was built using data collected from various states in India to train our model and test its accuracy in predicting the disease.

3.1 WHY THIS DATASET WAS CHOSEN ?

According to the knowledge gained from the review of literature and related work in this field , we can concur that dengue prevails in most of the regions in the Indian subcontinent and India , by itself , bearing a considerably higher percent of the world's dengue burden. On top of that , Dengue is an infection with high mortality rate and a public health concern in the country. All these factors helped to strongly conduct the analyses of the Indian population affected by Dengue and their typical symptoms. The amount of information acquired by the model about the disease and its traits has a direct impact on how accurate the resulting prediction would be.

3.2 ENVIRONMENT

The model was developed using Python programming language. We preferred Python over other programming languages as it is , by and large , greatly suitable for data analytics and machine learning tasks. The language is considered to be a consistent and concise programming platform with easy access to a diverse set of libraries with respect to machine learning such as pandas , numpy and scikit-learn.

The coding environment utilised in this project is Jupyter notebook as it computes the output immediately. As the project involves machine learning and few visualizations to support the work , we would need a lot of trial-and-error methods to check the proceedings then and there. Likewise , coding and output when viewed on the same screen , we gain confidence and learn new things about the working. The back-end kernel which performs all the working in the background . The best part is that coding can be done in cloud and hence , data can be kept safe and secure.

3.3 PHASE 1 : MACHINE LEARNING

Machine learning plays an important role in the healthcare industry. Algorithms can be worked with to predict the presence/absence of the disease (dengue) . Proper diagnosis/prognosis of the victim's symptoms stage wise can help doctors with important insights related to the disease. We made use of three machine learning algorithms in our project according to the complexity of our dataset and its attributes . Support Vector Machine (SVM) , Random Tree classifier and Logistic Regression were chosen to train our model with dengue data.

3.3.1 ROLE OF MACHINE LEARNING

A major role in machine learning goes to the classification step. Classification is the technique which shows us how the available data elements can be categorised. Random forest classifier was preferred over other classification techniques like Naive Bayes and decision trees as it is an ensemble algorithm and it addresses problems like overfitting extensively. It generates numerous decision trees , which in turn reduce the error rate of individuals and resulting in a more accurate result. Feature importance is a very beneficial step in this classifier. This helps us choose the better features or attributes that would give value to our model and improve the accuracy rate among the lot present. This is one of the main reasons of why we use Random forests to train our model . Furthermore , we can reduce the impact of overfitting on our model. The more the number of trees generated , the more stable performance of our classifier would be.

3.3.2 MACHINE LEARNING ALGORITHMS

SVM is used in cases when it is clearly evident that the data is non-linear and multi-dimensional and when drawing a straight line cannot classify the data. SVM separates the target data into two classes (in our case Dengue , positive or negative). Another technique known as Logistic Regression was implemented in our system which was determined to give a better prediction rate for dengue than SVM. Logistic regression is an appropriate technique when the dependent variable (target class) is binary. The working of this algorithm tends to describe the

relationship between the attributes present in the dataset and dependent variable. For example , when we talk about cardiovascular diseases , how the changes in blood pressure and cholesterol in our body together alter the probability of the person being affected by a heart disease. In our case , this regression technique answers the question of how the increase/decrease of temperature in a patient's body and presence of fatigue impact the likelihood of dengue in future. The predicting factors for dengue were studied to be fever , body temperature , muscle pain , eye pain , skin rash , fatigue and blood platelet count. The relationships among the variables are also described. The prediction rates are compared against each other to decide the better technique for our project domain. These estimates of accuracy aid us in designing a better model for the prediction of dengue disease with more finely tuned techniques favouring real time situations. The algorithms are discussed in detail in the next chapter.

3.4 PHASE 2 : IOT INTEGRATION

The second phase of the project is IoT integration in our environment. We will be simulating patient information in this phase. An IoT server connection is established and the data obtained about the patient is pushed to the cloud through the IoT platform. This will be greatly helpful during surveillance of the patient. Incase there is a patient with the probability of being affected by the disease and being monitored by a doctor , data can be entered to through the web and the system would tell the result (whether dengue is present or not). in the background , when the user enters the real data at that moment (physical factors) , the data is sent to the cloud and the predicted result is shown. The result is sent as an alert to the doctor as well if the condition is critical. The doctor will come to know the presence of disease without even visiting the patient. Adequate treatment or necessary steps can be advised to the patient accordingly. The alert message will contain information like GPS data (latitude and longitude) and the patient's condition. The alert text can be sent to any number given as an emergency contact. As the alert is sent as a text message , also there is no need for an active data connection. This system when integrated with an actual IoT processor like Arduino or Raspberry Pi , real time sensory information can be collected and uploaded to the cloud storage for the disease to be predicted.

3.4.1 IOT PLATFORM

Ubidots enables the user to upload data to a private cloud from any device connected to the Internet. Further actions are easily configurable after logging in to the account. This provides the user with a REST API to read/write data to the respective source. A personal API key is generated for each user account. TextLocal is used as the SMS gateway through which the alert message is

delivered. The system is integrated to the SMS API through web/application . The user's API key should be configured in the program.

The environment is built in Python programming language using Django , which is a web application framework which serves as a rapid programming/development framework for building web API's. There are various web development frameworks available with different project goals and needs. Django is one of the most common web frameworks used with Python. The main reasons to choose Django are that it maintains its own server , easy to code , secure , scalable , cross-platform and allows allows developers to reuse code.The user's application interface is accessed through this framework. We can develop a single project combining code for URLs , models , views and templates.

Chapter 4

Dengue Prediction System Design and Specifications

The implementation of the project dengue disease prediction is discussed in detail in this chapter. The author walks the readers through the projects's setup and working design , how the whole environment was developed from scratch , the technologies , techniques and tools used in the project to bring about a proper working model providing an acceptable result . The explanation for the methods applied given below and the results of evaluation detailed in the next chapter on the whole , would contribute to the overall understanding and precise interpretation of the reader.

4.1 DATASET :

The dataset used to train our dengue prediction model was gathered from actual medical data of patients containing the symptoms and physical changes experienced during the time of infection such as body temperature , eye pain and fever. These attributes as a whole are fed into our model so that the learning can be done by the system to predict the presence of dengue. The dataset was collected from Government hospitals and private hospitals . The dataset consisted of patient information from five Southern states of India namely Tamil Nadu , Andhra Pradesh , Telangana , Kerala and Karnataka which have a high mortality rate , high risk and epidemic record of the

disease. There are 16 attributes and 1000 records in the dataset. The description for the dataset is given below.

Gender - sex of the patient

Age - age of patient in years

Country - Nationality (India in this case)

State - name of state to which the patient belongs

Hospital - name of hospital which attended the patient

Year - Year when data of patients was generated

Temp - Average body temperature of the patient (in Fahrenheit)

Vomiting - if patient had feeling of nausea (Y / N)

Eye pain - patient faced pain in the eyes or not (Y / N)

Platelets - Number of platelets present in the patient’s body

Fatigue - if patient felt very tired(Y / N)

Fever - severity of fever(High / low)

Muscle pain - if patient experienced muscle pain (Y / N)

Skin rash - if patient had any kind of skin rashes (Y / N)

Disease - Target class (disease Positive or Negative)

Basically , the dataset provides possible information about the symptoms that could lead / have led to dengue in the patients recorded. The Disease attribute which showed the final result of whether dengue is present or not , is the dependent variable.

4.2 CODING ENVIRONMENT

The project was developed in Macintosh system running with macOS Mojave (10.14.3) , processing speed of 2.3 Ghz Intel Core i5 and 8GB RAM. The code to develop the project was written in Python programming language. Python v 3.7.3 was used. The Python 3.7 series is the latest major release of the Python language and it features several new features and optimisations. Hence , it provided a stable coding environment throughout the implementation of the project. Python is more about instinct and logic rather than syntax and code. Pandas can be used for general purpose data analytics , scipy and numpy to carry out advanced and complicated scientific and mathematical tasks whereas scikit-learn provides the actual machine learning coding.

The IPython Notebook also known as the Jupyter Notebook is an interactive environment powerful enough to combine code execution, plain text, mathematical functions , plots and visualizations and rich media support. It is a freely available , open-source notebook well known for its interactive and computational capability. It is an Integrated Development Environment (IDE) which can be used to develop code , instant output , text/ comments (to improve readability of the code) and various other resources. It serves as a best fit for most of machine learning , data science , exploratory data analysis . The Jupyter notebook version 5.7.8 was used for our project which can be navigated through Anaconda distribution . Anaconda is an open source environment that enables data analysts and datas scientists to develop and deploy machine learning techniques easily. The platform can manage a great many libraries and dependencies which compatible with Python packages (such as

scikit-learn and TensorFLow) specific to train machine learning and deep learning models. The first phase of the project was with machine learning. Here , all the necessary steps regarded to training and predicting the disease was performed.

4.3 COURSE OF ACTION

The breakdown for how all these steps were executed in this stage is given below:

- importing required libraries
- importing dataset
- Data preprocessing
- Splitting Dataset (Training and testing)
- Building and training model
- Testing for accuracy
- Prediction
- Visualisation of Data

4.3.1 : IMPORT REQUIRED LIBRARIES

The major libraries needed for developing our project are pandas , matplotlib , numpy and sci-kit learn. These are some built-in functions that come along with the Python installation. The libraries listed below are required to run the model without any interruption and are necessary for the working of the model.

- Pandas is a licensed Python library providing high performance data structures and tools for analysis of data. It creates a table object in memory known as DataFrame.
- Numpy is a core package in Python used for scientific computing. It enables enables faster processing of N-dimensional array objects . This library is basically used with numerical data.
- Matplotlib is a powerful library meant for plotting using Python and numpy. it is used to create data visualizations . It is used to create plots , graphs and figures like histograms , bar charts , pie charts , scatter plots , tables etc.
- Scikit-learn is the most productive and versatile library in Python for machine learning. It consists of many efficient tools that can help us achieve modelling and implement

machine learning including unsupervised and supervised techniques like classification , clustering and regression.

It is a better approach to explore the dataset in the first step before starting to work with it. The dataset file (dengue.csv) is imported into Jupyter , converted into a DataFrame structure in pandas and stored. The structure of the dataset can be viewed in Python using head() method which returns the top n rows (in default 5) of that particular dataframe.

4.3.2 DATA PREPROCESSING

Data Preprocessing is a cleaning technique which is used to convert / transform the raw data into a clean and properly structured dataset suitable for further analysis. Data is usually collected and gathered from various sources , so it should be good enough and in some specific format before the model learns or gets trained with the data. This will help us in achieving better and accurate results with valuable information. The basic steps in preprocessing involve filling up missing values and null values , getting rid of possible outliers and normalisation.

4.3.2.1 Label Encoding

The technique we follow to process our dataset is called Label encoding. There are generally multiple labels in some columns of a dataset which can be in word or number formats. Hence , to make such data more machine readable , we use Label encoder which converts categorical data into numeric values (between 0 and n-1) . It is very important to do this correctly in supervised learning methods like classification and regression.

We import the LabelEncoder() from sklearn's preprocessing package. The columns of vomiting , eye pain , skin rash , muscle pain , fatigue , fever and disease contain different categorical data. These are the values that should be encoded to numerical format so that our predictive model can better understand. Here , the values like high , low , yes , no , positive and negative are converted into numbers accordingly.

4.3.3 SPLITTING THE DATA

In machine learning , we usually split any dataset into two : training data and test data. The output variable along with other variables are included in the training set . The model learns the data and tries to generate some pattern . The other part of the dataset serves as a test set to validate our

model's prediction. The scikit library has a function called train_test_split to divide our data . test_size is the parameter which gives us the percentage of data that should belong to the test set. train_size stores the remaining part as the training dataset , either of which should be specified. random_state acts as a random number generator . For our dataset , we split the training and testing set with 80 , 20 ratio the random state is passed as 0.

4.3.4 TRAINING THE MODEL

As given earlier in the document , we will train our data with three machine learning techniques using Random Forest Classifier , SVM and Logistic Regression . The attributes we consider for training our model are basically the symptoms faced by dengue patients namely skin rash , muscle pain , fever , fatigue , blood platelet count , vomiting and temperature. The columns excluding the disease result are stored in a variable X. The variable showing the result is stored in Y. These are scraped out from the dataset using iloc method from pandas which is used in Python to select the specific rows and columns by indexing their respective locations. Given below are brief explanations of the working of algorithms used. The training set is stored in X_train and Y_train whereas the test set is saved in X_test and Y_test.

4.3.4.1 RANDOM FOREST CLASSIFIER

This algorithm operates as an ensemble by combining the power of individual decision trees , each predicting a result class out of the available data. This algorithm used divide-and-conquer approach. Though the lot can seem a bit uncorrelated , lacking mutual connection , together they give significant results. The error rates are also minimised by using this approach.

Random Forest algorithm plays an important role in featuring the importance of the attributes depending on what outcome we expect and the nature of our model. So , the predicted outcomes have very less interrelationship between each other. All the attributes we use to train our model have a significant impact on the final prediction of dengue , but the amount of impact they have on it makes the difference .

The random trees system is great at controlling issues like overfitting. Overfitting can be defined as a complication/ modelling error in machine learning techniques when a particular model

learns the given data too well and becomes excessively complicated. This issue can be addressed by techniques like cross validation.

4.3.4.2 SUPPORT VECTOR MACHINES

SVM is a great machine learning algorithm (supervised) highly preferred by machine learning enthusiasts for its accuracy rate with less computation and non-complex working. It can be considered a flexible algorithm for both classification and regression steps.

The SVM aims to detect a hyperplane in the multi-dimensional space with N being that the number of features/dimensions present. The task expected from SVM is to classify the data points into two separate classes. Many hyperplanes exist in the space to separate the points but the one with the highest margin should be found for the best results. Maximum margin is meant to be the longest distance between both the sets. An hyperplane is a subspace where dimension is one lesser. In other words , hyperplanes are decision boundaries chosen to divide the set apart. Support vectors are the influential factors among the data points which control the position and inclination of the hyperplane towards/away from the classifier margin.

$$c(x, y, f(x)) = \begin{cases} 0, & \text{if } y * f(x) \geq 1 \\ 1 - y * f(x), & \text{else} \end{cases}$$

The hinge loss function pushes towards maximising the margin. It is calculated if cost is 0 when prediction and originality are both either positive/negative. A regularisation parameter lends help to stabilise margin loss and maximisation.

$$\frac{\delta}{\delta w_k} \lambda \parallel w \parallel^2 = 2\lambda w_k$$

$$\frac{\delta}{\delta w_k} \left(1 - y_i \langle x_i, w \rangle\right)_+ = \begin{cases} 0, & \text{if } y_i \langle x_i, w \rangle \geq 1 \\ -y_i x_{ik}, & \text{else} \end{cases}$$

$$0 \leq h_\theta(x) \leq 1$$

$$w = w - \alpha \cdot (2\lambda w)$$

Partial derivatives decide the weight measures to find slope. When the model performs well and the predictions are accurate enough , the regularisation variable is adjusted , else both regularisation and loss variables are adjusted. (poor prediction)

$$w = w + \alpha \cdot (y_i \cdot x_i - 2\lambda w)$$

4.3.4.3 LOGISTIC REGRESSION

Logistic regression is a classification algorithm in machine learning which is suitable when the classification is binary (involving two class values). It is used for predictive analysis of data. It is very similar to linear regression model but it has its own characteristics. The dependent variable or outcome is continuous in nature and a straight line can connect both ends of the data in Linear regression. The regression type is logistic when the possible outcome is categorical / dichotomous , can have only limited values .(Yes / No , True / False , High/Medium/Low). It describes the association of of one dependent binary variable (disease) with nominal / ordinal set of independent variables. (symptoms of dengue). It is widely used to predict mortality and disease probability.

The Logistic Regression model uses a cost function known as Sigmoid function / logistic function . Hypothesis is that the cost function should be positive and range between 0 and 1 . The predictions are plotted against probable outcomes are dine using the Sigmoid.

$$f(x) = \frac{1}{1 + e^{-(x)}}$$

$$h\theta(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

The model can perform poorly if there is much noise and correlation in between the variables. Hence , it should be made sure to remove noisy data , outliers and closely related attributes. A better understanding can be reached when the results sections is read through along with appropriate results. The proceedings are discussed in the final chapter of the thesis.

Chapter 5

Implementation

5.1 RFC

The Random Forest classifier algorithm is imported from sklearn's ensemble.

A Gaussian classifier is created and stored in clf.

n_estimators is the parameter to specify number of trees in the forest. (1000 in our case)

The model is now trained using training data sets. (X_train , Y_train) which store the attribute values and disease result attribute respectively.

The model with randomised trees is built and trained using the dataset by .fit() method.

The prediction of this algorithm is done using .predict() method and the X_test variable is passed as the parameter. As soon as the model runs , it displays a range of parameters such as max_depth , min_samples_split , min_weight_fraction_leaf which are results of background working of the algorithm . The feature importance of the attributes used in our model are calculated using feature_importances_ method. The ranking of attributes with respect to their importance for the

model is found. Pruning is a technique used in random forest model to limit the number of trees and create a more effective subset of trees and reduce the chances of overfitting. With the result of important features , another random model was built to compare the accuracy of both models with all attributes and important attributes. to quantify the quality of our predictions , we will use the AUC (Area under the ROC curve) which is a metric in Python that will give us the aggregate performance measure of our model . This value ranges between 0 and 1 depending on how well the model has classified the data. In our case , more the AUC score , more the ability of the model to diagnose patients with / without the disease.

5.2 SVM

We defined our SVM model by importing the svm module from sklearn package , which contains all the default classes for SVM algorithms. As our task is to classify the data , we will use Support Vector Classifier (SVC). An important parameter for this class is kernel type which depends on the linearity of the data. The kernel type is set to linear in our model. C is the regularisation parameter which tells the SVM about misclassification of data while training. It is a trade-off measure between low testing error and training error. Higher the C value , lower the margin in SVM. It is set to 1 in our model. The model is fit using training data (X_train , Y_train) by fit() method and predictions are generated using predict() method. Similar to the previous model , roc-auc score is estimated to check the accuracy. We also use other measures like Precision and Recall which evaluate the output quality of there classification performed. They're inversely proportional measures. These metrics can be calculated by importing recall_score and precision_score from sklearn.metrics package.

5.3 LOGISTIC REGRESSION

The logistic regression model for our data was built by importing the logistic regression module and creating a logistic classifier object (regressor). The attributes to be fed into the model are selected from the dataset and stored in X and Y variables in its initial stage The predicted scored for logistic model are calculated . We also use a confusing matrix here to further rate our model in terms of accuracy. It returns the number of correct and incorrect values according to both the classes.

5.4 INTERNET OF THINGS (UBIDOTS AND TEXTLOCAL)

The Ubidots Python API Client connects to the Ubidots API. The Ubidots package is installed in the system from the command line using pip , python package installer. The Ubidots version used for our project is Ubidots 1.6.6. We can generate a default API token or personal key in the Ubidots account. This API token is used to establish connection with the API. the APiClient is imported into the system from Ubidots package. An instance of the APIClient is created with the API token stored in it. Variables retrieved from the website dashboard are instantiated in the Python environment with the respective values.

The machine learning model is now integrated into the Ubidots platform by using input() method which fetches the user input for the model. In our system , the user enters the different values of attributes (symptoms) in the input text box for the model to work out and predict if the disease is present or not. The input method is used to input multiple values from the user in the same line separated by white spaces. The split() breaks the string using white spaces. The map() applies the regression function and returns a list of results.

The packages necessary for setting up the IoT platform and SMS gateway are imported such as requests , json and urllib. Requests enable the user to send various types of HTTP requests and access response data in Python. Json is a default package in the python library which is used to encode JSON format into bytes which are to meant to be transmitted over our network. The json.loads() convert the JSON encoded data into Python instances. ; geo_req.text fetches the internal GPS location . A user-defined method sendsms is created A response variable is created with the parameters that are used to send text alerts depending on the condition.

TextLocal is an online platform to integrate SMS API with our system. The SMS API key generated from the user's TextLocal account is stored in apikey , the respective emergency contact is specified in numbers , sender is passed as TXTLCL. The username of the sender and the message content are passed along with the former parameters in the sendsms method. urllib.request and urllib.parse are imported from Python which are used to fetch URLs (Uniform Resource Locator) and authentication purposes and manipulating URL by encoding the strings given as input (username , sender , message) into separate components. The character encoding format used is UTF-8 which is specified in data.encode().

The real time geolocation data of the user is generated from ipstack.com which maps the user's IP address to the country , region , latitude/longitude and city to which the user belongs. Another API access key which is a unique authentication for the user is appended to our code which establishes connection with the base URL.

send_url=http://api.ipstack.com/37.228.249.171access_key=c891ac9ebd9a28f43c2c32fd34e4298c

The front end user interface is coded in Django within a virtual environment. The respective Python code is written in the Python files (manage.py , settings.py , urls.py) and web interface for the user is designed in HTML. Python and HTML merged together produce a web framework for the user to know if the disease is positive or negative , depending on the inputs given by the same. These values are given as data to the best disease prediction model. The results are discussed in the final chapter.

Chapter 6

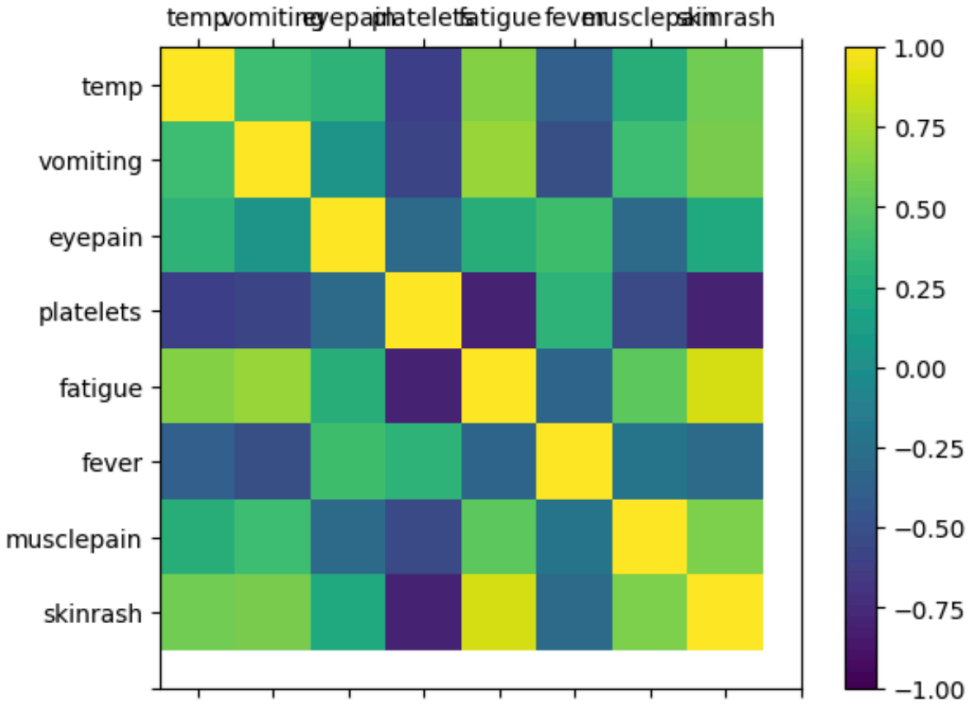
Results and Visualizations.

The results we arrived at from the three models are spoken about in detail here in this chapter. These are the major factors which help us decide which machine learning technique would be feasible and accurate to develop a model to predict the existence of dengue fever.

6.1 RANDOM FOREST CLASSIFIER

When the Random forest model was trained using the dengue symptoms data , a classifier object was created and stored in clf. Certain parameter were returned after the training stage such as max_depth (maximum depth of tree) , n_jobs (no.of jobs for fit and predict) , min_samples_split (no.of samples needed to split the nodes). These outputs show us that the classifier model was successfully built. The testing data (X_test) was used to evaluate the prediction our classifier model returned. The predicted results were stored in a variable Y_pred.

After the necessary imports were done , accuracy score of our RFC model was estimated by comparing the disease values from testing set and predicted set. The accuracy was noted to be 98%. The reason to use random forest model was to measure the feature importance of the attributes used rather than to test the accuracy. The column labels were stored in columns variable as a series of strings. The pandas dataframe was stored in a new variable df_1. The feature importance values were returned in descending order from highest rank to lowest rank. It was pretty much evident from these results that these symptoms have a good amount of impact on the disease. Our results supported the previous findings as well. The blood platelet count of the patient with an importance score of 0.361825 influenced the disease to a great degree followed by fatigue , skin rash , temperature , vomiting , muscle pain , eyes pain and fever.



CORRELATION OF VARIABLES

The correlation plot of the influential attributes from the dataset are generated and plotted using dataframe.corr() method which helps us to study the internal relationship between the various variables. The value ranges between -1 to +1 through 0 meaning negative , positive and no correlation . The statistical relationship between the variables are noted by the intensity of the color in the plot.

6.2 SUPPORT VECTOR MACHINE

The SVM model was built and the classifier object was returned. roc_auc_score where the model's performance was measured from the testing set (Y_test) and predicted values(predictions). SVM gave an accuracy rate of 94.15% compared with the actual result set. The hypothesis is that logistic regression would be more suitable and feasible for the prediction of dengue fever than the support vector model. The regressor model was trained using logistic regression and predicted results were stored in Y_pred.

A confusion matrix was drawn after training the model , which is a table describing the performance of any classifier model having known the results in advance. (i.e how to model had confused itself between the predictions and true values) This was calculated by using confusion_matrix() method.

	Class 1 Predicted	*Class 2 Predicted*
Class 1 Actual	**TP**	**FN**
Class 2 Actual	**FP**	**TN**

CONFUSION MATRIX

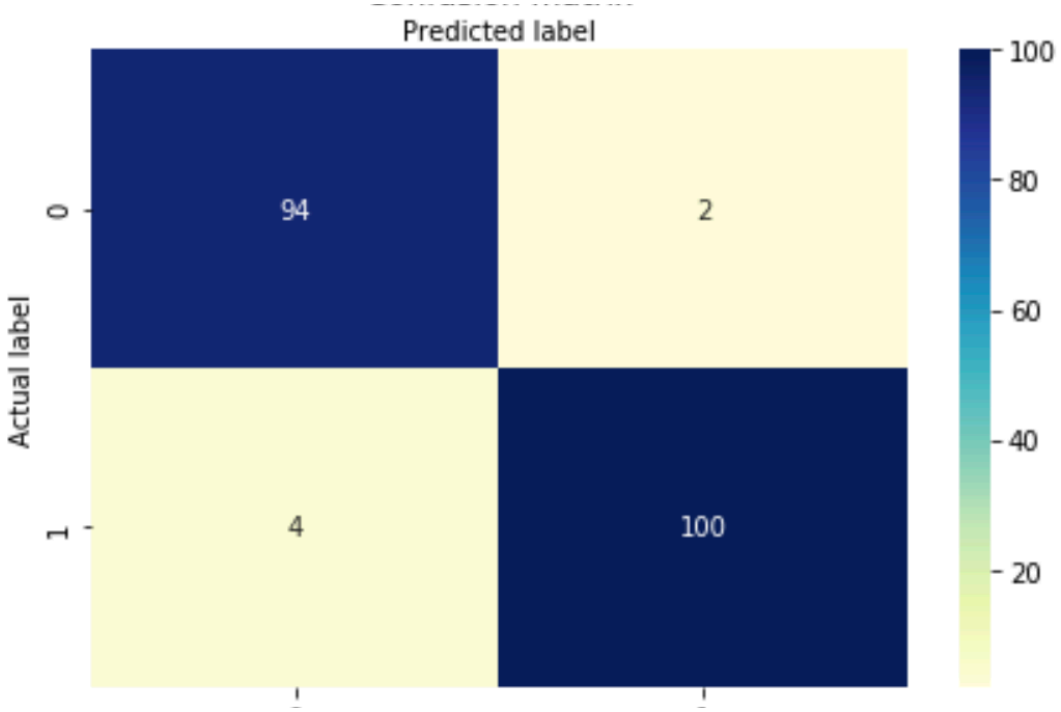
True Positive (TP) : Observation and prediction are both positive.

False Negative (FN) : Observation positive but prediction false.

True Negative (TN) : Both observations and predictions are negative.

False Positive (FP) : Prediction positive but observations are false.

The confusion matrix was visualised using a heat map. The findings were , out of 200 prediction made to guess the disease , the SVM model predicted disease positive 102 times and



CONFUSION MATRIX : HEAT MAP

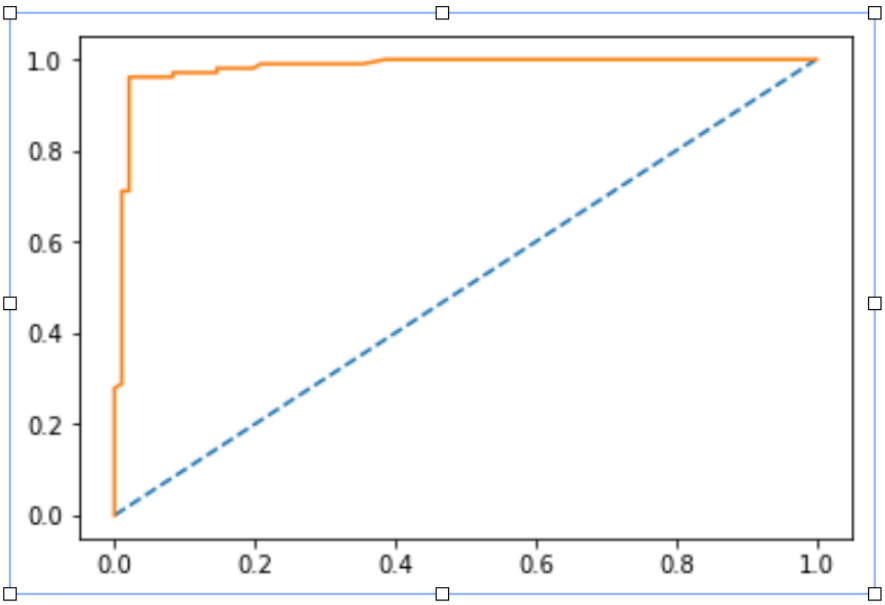
disease negative 98 times.

6.3 LOGISTIC REGRESSION

Logistic regression was found to be more accurate than SVM model in predicting the disease. The auc score calculated for the model was 97.03% which is an excellent result . This proves that the logistic model is a better way to predict dengue in a patient. Precision and Recall were also calculated for the logistic model and then values were 97.9% and 90% respectively. Precision is the exactness measure which tells us how relevant the model had predicted the values whereas recall shows us how many relevant predictions the algorithm returned. The Roc curve (receiver operating characteristics) is a graph which describes the performance of a classification model at different thresholds. The True positives and False Negatives are plotted against each other.

6.3.1 ROC-AUC

The AUC (Area under ROC curve) explains more about the correctness of the model's prediction. AUC ranges from 0 to 1 denoting the accuracy rate of the model. More the curve tends



ROC CURVE

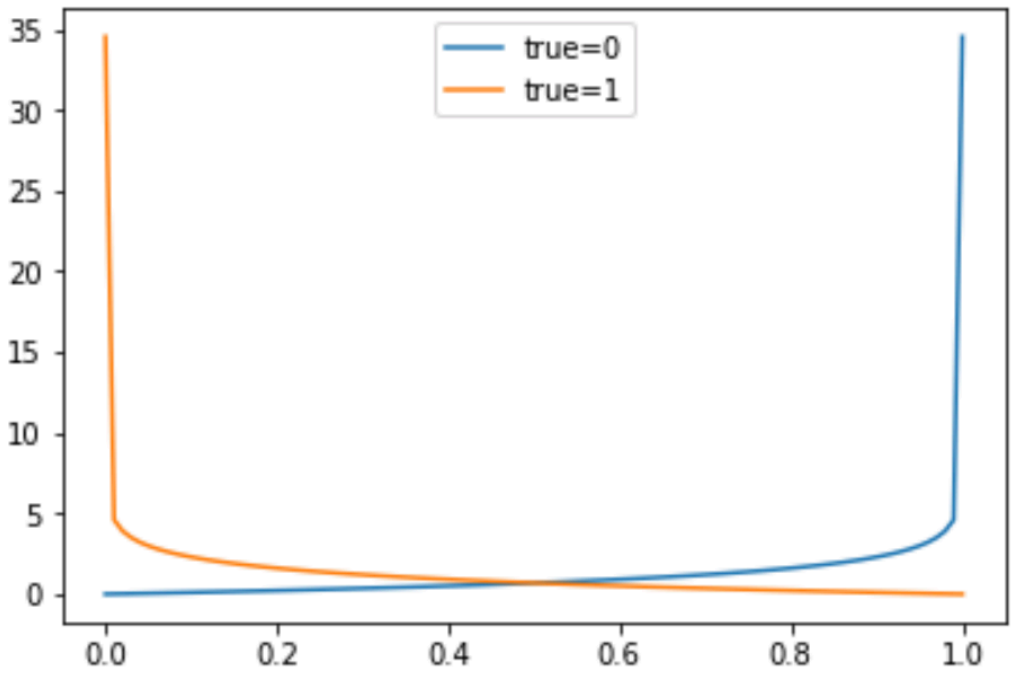
towards 1 , better the accuracy of the model. These metrics also talk about th specificity and sensitivity of the model. If we want to consider positive results , sensitivity is important and if we want the to identify negative results , specificity should be taken.

Here the curve tends more towards the top-right corner of the graph which means the overall accuracy of the model is good. Hence , we can say that higher the score , better the model is at predicting the disease.

6.3.2 LOG LOSS

A log loss curve is plotted for the model as well as it is a binary classifier. It is a goofd metric when logistic regression model is implemented. The AUC is calculated to test the classification of the binary result variable , whereas the log loss gives the certainty of the classification. Log loss is the measure of how uncertain the model predicts compared to the actual results. A log loss value of 0 means the model is at its best. The aim of the user is to minimise this score as much as possible to produce a good model. The log loss value for our model is 0.166. The visualisation is given below.

The graph below describes the Log Loss contribution from a positive instance where the predicted probability ranges from 0 (totally wrong prediction) to 1 (exact prediction). This shows that the model has predicted the disease very well.



LOG LOSS CURVE

Chapter 7

Conclusion

From the experience of training dengue prediction models using various machine learning algorithms , we gain ample knowledge about the working and implementation of them. This helps us to apply the models to appropriate real world situations. Machine learning and Internet Of Things are saviours in the field of medicine and healthcare. The power of prediction of diseases and the ubiquity of computing , storage and data enabled by IoT will definitely be the future in various medical scenarios. With respect to Dengue , we have now been through the working and demonstration of three machine learning algorithms . Random Forest Classifier described the main factors that would impact the dengue infection in humans namely blood platelet , fatigue and skin rash. The Support vectors generated a good accuracy of 94.15 % but the logistic regression performed better in our case with a prediction rate of 97.03% . After all , we can decide that logistic regression technique would be the best suitable algorithm for dengue prediction. Besides , the event of user knowing the result of whether the infection is present or not using a personal web interface adds more value to the project. The SMS alerts can be sent to emergency contact numbers such as a doctor or family members. Further steps or treatment can be advised to the patient by studying the input factors given by the user (symptoms) and even the user can be located with the help of GPS information collected. This would be a really powerful model design if sensory information could be collected from IoT sensors around the user's living space. Future research should be enforced to eradicate or reduce the risk of dengue in India as well as the world. Enhanced treatments and Vaccinations should be produced as the burden of dengue is increasing rapidly throughout the globe.

Chapter 8

Bibliography

1 . Gubler, D. (1998). Dengue and Dengue Hemorrhagic Fever. Clinical Microbiology Reviews, 11(3), pp.480-496.

2. Gubler, D. and Clark, G. (1995). Dengue/Dengue Hemorrhagic Fever: The Emergence of a Global Health Problem. Emerging Infectious Diseases, 1(2), pp.55-57.

3.Henchal, E. and Putnak, J. (1990). The dengue viruses. Clinical Microbiology Reviews, 3(4), pp. 376-396.

4. Thein, S., Aaskov, J., Shwe, T., Aye, K., Aung, M., Zaw, A., Aye, K. and Aye, M. (1997). Risk Factors in Dengue Shock Syndrome. The American Journal of Tropical Medicine and Hygiene, 56(5), pp.566-572.

5.Gibbons, R. (2002). Dengue: an escalating problem. BMJ, 324(7353), pp.1563-1566.

6. Guzman, M., Kouri, G., Bravo, J., Soler, M. and Martinez, E. (1991). Sequential infection as risk factor for dengue hemorrhagic fever/dengue shock syndrome (DHF/DSS) during the 1981 dengue hemorrhagic cuban epidemic. Memórias do Instituto Oswaldo Cruz, 86(3), pp.367-367.

7. Bhatt, S., Gething, P., Brady, O. and Messina, J. (2013). The global distribution and burden of dengue. Nature, 496(7446), pp.504-507.

8.HOLMES, E. and TWIDDY, S. (2003). The origin, emergence and evolutionary genetics of dengue virus. Infection, Genetics and Evolution, 3(1), pp.19-28.

9.Gaunt, M., Sall, A., Lamballerie, X., Gould, E., Falconar, A. and Dzhivanian, T. (2001). Phylogenetic relationships of flaviviruses correlate with their epidemiology, disease association and biogeography. Journal of General Virology, 82(8), pp.1867-1876.

10.Dar, L., Broor, S., Sengupta, S. and Seth, P. (1999). The First Major Outbreak of Dengue Hemorrhagic Fever in Delhi, India. Emerging Infectious Diseases, 5(4), pp.589-590.

11. Chaturvedi, U. and Nagar, R. (2008). Dengue and dengue haemorrhagic fever: Indian perspective. Journal of Biosciences, 33(4), pp.429-441.

12.Organization, W. (2019). Dengue guidelines for diagnosis, treatment, prevention and control : new edition. [online] Apps.who.int. Available at: https://apps.who.int/iris/handle/10665/44188 [Accessed 2 Sep. 2019].

13.Organization, W. (2019). Dengue haemorrhagic fever : diagnosis, treatment, prevention and control. [online] Apps.who.int. Available at: https://apps.who.int/iris/handle/10665/41988 [Accessed 2 Sep. 2019].

14.Global strategic framework for integrated vector management. (2004). Geneva: World Health Organization.

15.H. M. Aburas, B. G. Cetiner, and M. Sari, "Dengue confirmed-cases prediction: A neural network model," Expert Systems with Applications, vol. 37, no. 6, pp. 4256–4260, 2010.

16.M. Akshaiya, M. Akshaya, G. A. Lakshmi, and S. Sumathi, "A GPS Based Dengue Risk Index to Predict the Susceptibility of an Individual," 2018 International Conference on Communication, Computing and Internet of Things (IC3IoT), 2018.

17. S. Gambhir, Y. Kumar, S. Malik, G. Yadav, and A. Malik, "Early Diagnostics Model for Dengue Disease Using Decision Tree-Based Approaches," Pre-Screening Systems for Early Disease Prediction, Detection, and Prevention Advances in Medical Diagnosis, Treatment, and Care, pp. 69–87.

18.P. Guo, T. Liu, Q. Zhang, L. Wang, J. Xiao, Q. Zhang, G. Luo, Z. Li, J. He, Y. Zhang, and W. Ma, "Developing a dengue forecast model using machine learning: A case study in China," PLOS Neglected Tropical Diseases, vol. 11, no. 10, 2017.

19.F. Ibrahim, M. N. Taib, W. A. B. W. Abas, C. C. Guan, and S. Sulaiman, "A novel dengue fever (DF) and dengue haemorrhagic fever (DHF) analysis using artificial neural network (ANN)," Computer Methods and Programs in Biomedicine, vol. 79, no. 3, pp. 273–281, 2005.

20.R. Kapoor, J. S. Sidhu, and S. Chander, "IoT based National Healthcare Framework for Vector-Borne diseases in India perspective: A Feasibility Study," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018.

21.M. Kassim, N. A. N. Ali, A. Idris, S. Shahbudin, and R. A. Rahman, "Dengue Attack Analysis System on Mobile Application," 2018 IEEE 8th International Conference on System Engineering and Technology (ICSET), 2018.

22.J. Ong, X. Liu, J. Rajarethinam, G. Yap, D. Ho, and L. C. Ng, "A novel entomological index, Aedes aegypti Breeding Percentage, reveals the geographical spread of the dengue vector in

Singapore and serves as a spatial risk indicator for dengue," Parasites & Vectors, vol. 12, no. 1, Aug. 2019.

23 S. Polwiang, "The correlation of climate factors on dengue transmission in urban area: Bangkok and Singapore cases," 2016.

24 S. Polwiang, "The correlation of climate factors on dengue transmission in urban area: Bangkok and Singapore cases," 2016.

25 P. Siriyasatien, S. Chadsuthi, K. Jampachaisri, and K. Kesorn, "Dengue Epidemics Prediction: A Survey of the State-of-the-Art Based on Data Science Processes," IEEE Access, vol. 6, pp. 53757–53795, 2018.

26H. Somwanshi and P. Ganjewar, "Real-Time Dengue Prediction Using Naive Bayes Predicator in the IoT," 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), 2018.

27 S. Swain, M. Bhatt, S. Pati, and R. J. S. Magalhaes, "Distribution of and associated factors for dengue burden in the state of Odisha, India during 2010–2016," Infectious Diseases of Poverty, vol. 8, no. 1, Jun. 2019.

28 K. S. Vannice, A. Durbin, and J. Hombach, "Status of vaccine research and development of vaccines for dengue," Vaccine, vol. 34, no. 26, pp. 2934–2938, 2016.