**DENGUE DISEASE PREDICTION & ANALYSIS**

**USING MACHINE LEARNING**

**A PROJECT REPORT**

*Submitted by*

**VAIBHAV DADHICH [Reg No: RA1611008010189]**

**MOHNISH KHERA [Reg No: RA1611008010285]**

**MANU ANAND [Reg No: RA1611008010441]**

**PRANAV JHA [Reg No: RA1611008010677]**

*Under the Guidance of*

**MRS. G. GEETHA**

(Assistant Professor, Department of Information Technology)

*In partial fulfillment of the Requirements for the Degree*

*of*

**BACHELOR OF TECHNOLOGY**



**DEPARTMENT OF INFORMATION TECHNOLOGY**

**FACULTY OF ENGINEERING AND TECHNOLOGY**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

**KATTANKULATHUR - 603203**

**KATTANKULATHUR – 603203**

**APRIL 2020**

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

KATTANKULATHUR-603203

BONAFIDE CERTIFICATE

Certified that this project report titled “**DENGUE DISEASE PREDICTION & ANALYSIS”** is the bonafide work of **“VAIBHAV DADHICH [Reg No: RA1611008010189], MOHNISH KHERA [Reg No: RA1611008010289], MANU ANAND [Reg No: RA1611008010441], PRANAV JHA [Reg No: RA1611008010677],** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion for this or any other candidate.

|  |  |
| --- | --- |
| MRS. G. GEETHA  GUIDE  Assistant Professor  Dept. of Information Technology | Dr. G. VADIVU HEAD OF THE DEPARTMENT Dept. of Information Technology |
|  |  |
| Signature of Internal Examiner | Signature of Internal Examiner |

**ABSTRACT**

Dengue is a scourged infection found in a few developed just as developing nations like India. Late research evaluates that there was 390 million dengue contamination for each year and predicts that dengue transmission is available all through the tropics, with the most elevated hazard in the American and the Asian locales. This infection is brought about by reproducing a mosquito called Aedes. Prolonged rainstorm downpours give progressively dormant water to mosquitoes that convey these infections to breed. Across India, 70 individuals have been a target of this deadly disease and more than 36,000 individuals have been influenced by dengue since January, as indicated by various surveys. Dengue fever influences the body by causing a high fever and influenza-like indications. The infection contaminates your blood and causes cerebral pain, rash, itchiness, squeamishness, tiredness, and fever. Climatic and socioeconomic variables impact the dengue ailment. Because of the adjustments in climate conditions like temperature, humidity, wind heading and precipitation prompts the skewering of dengue infection. Dengue has become a worldwide issue and it is regular in all the nations. In this project, we look at different algorithms of machine learning for example (SVM) Support Vector Machines, Decision Tree and Tree Bagger and found the precision of these calculations for the expectation of dengue ailment. Data Mining is the system for the grouping or classification of sicknesses like dengue. Right now, is utilized to assess and analyze the outcomes. MATLAB software is used to compile and run these algorithms to compare the results and find out the best one.

**ACKNOWLEDGEMENT**

The success and the outcome of this project required guidance and assistance from different sources and we feel extremely fortunate to have got this all along with the completion of our project. Whatever we have done is largely due to such guidance and assistance and we would not forget to pay our sincere thanks to them.

We owe our profound gratitude to the Head of the Department, Department of Information Technology, **Dr. G. Vadivu**, for all the help and infrastructure provided to us to complete

this project successfully and her valuable guidance.

We express our sincere thanks to our project guide **Ms. G. Geetha**, who took a keen interest in our project work and guided us all along, till the completion of our project work by providing all the necessary guidance for developing a good system.

We are thankful for and fortunate enough to get constant encouragement, support and guidance from all the teaching staff of the Department of Information Technology which helped us in completing our major project work. Also, we would like to extend our sincere regards to all the non-teaching staff of the Department of Information Technology for their timely support.

**Vaibhav Dadhich RA1611008010189**

**Mohnish Khera RA1611008010285**

**Manu Anand RA1611008010441**

**Pranav Jha RA1611008010677**

**TABLE OF CONTENTS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| **CHAPTER NO. TITLE** | | |  | **PAGENO.** |
|  | ABSTRACT | |  | iii |
|  | ACKNOWLEDGEMENT | |  | iv |
|  | LIST OF TABLES | |  | vi |
|  | LIST OF FIGURES | |  | vii |
|  | LIST OF SYMBOLS, ABBREVIATIONS | | | viii |
|  |  | | |  |
| 1. | INTRODUCTION | |  | 1 |
|  | 1.1 | GENERAL |  | 2 |
|  | 1.2 | PURPOSE | | 2 |
|  | 1.3 | MOTIVATION AND BACKGROUND |  | 3 |
|  | 1.4 | PROBLEM STATEMENT |  | 4 |
|  | 1.5 | RESEARCH OBJECTIVE |  | 4 |
|  | 1.6 | REPORT ORGANIZATION | | 5 |
|  | 1.7 | EXISTING SYSTEMS | | 5 |
|  | 1.8 | ISSUES IN EXISTING SYSTEM | | 6 |
| 2. | LITERATURE SURVEY | |  | 7 |
| 3. | SYSTEM ANALYSIS | | | 11 |
| 4. | SYSTEM DESIGN | | | 13 |
| 5. | PROPOSED METHODOLOGY | | | 18 |
| 6. | SYSTEM TESTING | | | 22 |
| 7. | EXPERIMENTAL RESULTS | | | 29 |
|  | 7.1 COMPARISON GRAPHS | | | 30 |
|  | 7.2 OUTPUT SNAPSHOTS | | | 33 |
| 8. | CONCLUSION | | | 36 |
| 9. | FUTURE ENHANCEMENTS | | | 38 |
| 10. | REFERENCES | ENCES | | 39 |
|  | APPENDIX | NDIX | | 41 |
|  | PAPER PUBLICATION STATUS | R PUBLISH STATUS | | 52 |
|  | PLA | GIARISM REPORT | | 53 |
|  |  |  | |  |

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
| S.No | Table | Page |
| 1 | Different Testcases with their obtained outcomes | 23 |
| 2 | Confusion Matrix | 24 |
| 3 | Quantitative Evaluation with different models | 31 |

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| S. No. | Name | Page No. |
| Fig. - 1 | System Architecture | 24 |
| Fig. - 2 | Selection of algorithms | 25 |
| Fig. - 3 | Use-case Diagram | 26 |
| Fig. - 4 | Class diagram | 27 |
| Fig. - 5 | Understanding the data set | 30 |
| Fig. - 6 | Training data set | 31 |
| Fig. - 7 | Testing data set | 31 |
| Fig. - 8 | Confusion Matrix | 33 |
| Fig. - 9 | Working of Decision Trees | 35 |
| Fig. - 10 | ROC Curve Analysis vs Machine Learning Models | 42 |
| Fig. - 11 | Graphical Analysis of Accuracy | 43 |
| Fig. - 12 | Graphical Analysis of Precision | 43 |
| Fig. – 13 | Graphical Analysis of Recall | 45 |
| Fig. – 14 | Prediction data set | 46 |
| Fig. – 15 | Training Set | 46 |
| Fig. – 16 | Testing Set | 47 |
| Fig. – 17 | Prediction result based on SVM | 47 |
| Fig. – 18 | Confusion Matrix of SVM | 48 |
| Fig. – 19 | Evaluation Metrics of SVM | 48 |
| Fig. – 23 | Evaluation Metrics of DT | 49 |
| Fig. – 26 | Evaluation Metrics of TB | 49 |
| Fig. – 28 | Misclassification curve | 50 |

**ABBREVIATIONS**

**AI:** Artificial Intelligence

**CPS:** Cyber-Physical System

**GUI:**  Graphical User Interface

**ANN:** Artificial Neural Networks

**SVM:** Support Vector Machine

**DT:** Decision Tree

**TB:** Tree Bagger

**LIST OF SYMBOLS**

|  |  |
| --- | --- |
|  | Summation |
|  | Alpha |
| **%** | Percentage |
|  | Lambda |

**CHAPTER 1**

**INTRODUCTION**

Dengue fever is an overpowering ailment in like manner called a break bore fever or dandy fever or dengue hemorrhagic fever brought about by a gathering of an infection that is transmitted by mosquitoes. The group of mosquitoes that transmits dengue infection will likewise prompt ailments like chikungunya, yellow fever and Zika infection. Zika and dengue diseases are named organic wellbeing level (BSL) 2 pathogens while the chikungunya infection is named a BSL-3 operator. The side effects of Zika are similar to those of dengue and chikungunya, which are diseases brought about by different infections spread by a similar kind of mosquitoes. The infection is requested into four serotypes insinuated as DV-1, DV-2, DV-3 and DV-4, where DV is a positive-abandoned embodied RNA contamination. It is made out of three fundamental protein characteristics, which encode the nucleocapsid or center (C) protein, a layer related (M) protein, an envelope (E) glycoprotein, and seven non-auxiliary (NS) proteins. As per the WHO report, in 2015, India recorded a serious episode since 2006 with in excess of 15000 cases. Americans revealed more than 2.38 million cases in 2016 and 50172 instances of dengue fever in 2017, a decrease as contrasted and earlier years, it assesses that 50 to 100 million contamination happen early which incorporate 5 lakh dengue fever cases and 22,000 passing, for the most part among kids. An infection can impact anyone yet is progressively outrageous in immunodeficiency people. This tainting is brought about by the eat of Aedes aegypti, a female mosquito. Exchange mosquitoes get sullied because of the affirmation of the blood from a man who is spoiled with dengue disease.

**1.1** **GENERAL**

The design of the input is the connection between the system and the client. It involves the creating detail and strategies for information planning and those means are important to place exchange information into a usable structure for preparing can be accomplished by investigating the PC to peruse information from a composed or printed report or it can happen by having individuals entering the information legitimately into the framework. The plan of information centers around controlling the measure of info required, controlling the blunders, maintaining a strategic distance from delay, dodging additional means and keeping the procedure basic. The info is structured in such a manner in this way, that it furnishes security and usability with holding protection. Information Design thought about the accompanying things:

1. What information ought to be given as input?
2. How the information ought to be organized or coded?
3. The exchange to control the working faculty in giving input.
4. Methods for getting ready information approvals and steps to follow when a mistake.

**1.2** **PURPOSE**

Information Design is the way toward changing over a client arranged portrayal of the contribution to a PC based framework. This structure is essential to stay away from blunders in the information input procedure and demonstrate the right course to the administration for getting right data from the modernized framework.

It is accomplished by making easy to understand screens for the information passage to deal with an enormous volume of information. The objective of structuring input is to make information section simpler and to be liberated from mistakes. The information passage screen is structured so that all the information controls can be performed. It likewise gives record seeing offices.

When the information is entered it will check for its legitimacy. Info can be entered with the help of screens. Proper messages are given with respect to when required so the client won't be in the maize of moment. Subsequently, the goal of information configuration is to make an info format that is user friendly.

**1.2.1** **SCOPE**

The scope of this model is very vast. During the entire time of the year, every health care facility wastes many man-hours on the analysis for the accuracy, precision of the diagnosis using previously recorded data from past years. This prediction system will act as a savior for those health care infrastructures as they can utilize their time in other activities.

For example, if Hospital A uses this prediction system all across their health care infrastructures, it can help them to utilize their crucial time performing better tests and improving their testing facilities. The scalability of this project is immense as it can be used on a large scale to benefit the health care systems exponentially. The features and parameters of this project are configurable and can be scaled very easily as well.

**1.3 MOTIVATION AND BACKGROUND**

We gathered our collective motivation to research on this topic and do our homework after getting in touch and taking an in-depth site from our co-students in the medical department.  
Also, considering the fact of how deadly and lethal this infection can be provided it doesn't discriminate between its victim. The platelet count can go very low that can result in severe medical issues and a long-lasting recovery if proper medical aid is not provided.  
Even though this is curable to a large extent, the number of cases per year in our country has not shown any significant reduction over the years.

This calls for the efficient use of technology, Machine Learning in our case, to reduce the time and efforts to cure this condition as much as possible. Even though the disease is not concentrated in any specific region in the world, getting affected by this anywhere across the globe is painful and can hinder our day to day exercises. Having attended a workshop on Dengue Disease & it’s complications by AIIMS, Delhi, that was enough a motivation for us to pursue this topic.

**1.4** **PROBLEM STATEMENT**

Implement various Machine Learning algorithms and perform Data Analysis to predict the accuracy and precision of the diagnosis.

This will be achieved by performing the following:

1. Obtain and gather raw data.
2. Pre-process these raw data sets.
3. Perform data visualization and data analysis.
4. Choose the various algorithm to be implemented
5. Perform feature selection to enhance the accuracy
6. Execute algorithms and obtain accurate, precise verification.
7. Final analysis and documentation

**1.5 RESEARCH OBJECTIVE**

The primary objective of our research is the health care facility and how we can help them in diagnosing Dengue per se, more accurately and precisely. Our major objective is to help the health care facilities prepare in advance for a more precise and accurate diagnosis. The last thing any patient would want is being treated for some other disease whilst having another.

Advantage of our research work is as below:

1. Better diagnosis of Dengue patients.
2. Time-saving approach since it’s processed by computers.
3. Improved health care facility and patient relationships.
4. Help the health care facility in this intense competition of providing better medical aid.
5. Enhanced employment creation and its management.
6. More precise and accurate diagnosis means better medical assistance to the patients.

**1.6 REPORT ORGANIZATION**

This report showcases the study we have done sequentially. In Chapter - 1, we have discussed Dengue Disease and its impacts. In the following Chapter - 2 we have briefed on the previous work and study of researchers on our emphasized disease: Dengue. We have discussed their work and come to conclusions about this widespread infection.

Then we proposed methodology as part of Chapter - 3, where we first describe our data set and then started the process flow of our approach of studying the data. We have also specified the requirements to carry out the study for the same.

In Chapter - 4, we progress towards data visualization, where we study the major aspect of data and graphically represent our conclusions and study. It's the data exploration phase. Following that is our Chapter - 5, where we give details of models and algorithm in depth that we have used to determine the accuracy and precision of the data. We have showcased our conclusions, output and results in Chapter - 5, and chosen our best model.

In Chapter - 6 and have discussed the future scope of our study in. Ending with links and reference to papers, study sources we have included our code for models in Appendix-I and code for visualization in Appendix-II.

**1.7** **EXISTING SYSTEM**

In existing, they used naïve Baye's count and moreover researches what rule can act right now the future desire. The objective is to settle on an estimate model by using choice tree for anticipating the chances of occasions of dengue ailments in an inborn region. The database is examined for the creation of a solo model to envision the chances of hitting the sickness using the regulated classifier model.

**1.8 ISSUES IN EXISTING SYSTEM**

The existing scenario is based on naïve Baye’s model which provides less prediction accuracy. When Accuracy, Precision and Recall are calculated on a collective bases then the algorithm provides very less prediction accuracy as compared to the algorithms used in the proposed system.

The execution time is high. Nowadays, in most of the machine learning algorithms, all the time taken is mostly for training and testing of data, after that, execution time is very less as compared to the existing system which is using naïve Baye’s model.

**CHAPTER 2**

**LITERATURE SURVEY**

**2.1 IDENTIFYING DENGUE DISEASE IN MEDICAL TERMS**

1. The global distribution of dengue [1]. Dengue is a foundational viral disease transmitted between people by Aedes mosquitoes. For certain patients, dengue is a perilous disease. Presently, we lack approved antibodies, express therapeutics, and huge vector control attempts have not stopped its fast turn of events and overall spread. The contemporary overall dispersion of the danger of dengue infection and its general wellbeing trouble are ineffectively known. Here we attempt a thorough get together of known records of dengue event around the world, and utilize a proper displaying structure to outline worldwide appropriation of dengue chance. We at that point pair the subsequent hazard map with gritty longitudinal data from dengue accomplice studies and populace surfaces to construe the general wellbeing weight of dengue in 2010.
2. Daily forecast of dengue fever incidents [2]. This paper proposed a continuous dengue hazard forecast for a little zone. They utilized hazard forecast models rather than a customary measurable model for early notice, target reconnaissance, and mediation. The accuracy of the spatial and worldly units can be effortlessly acclimated to various settings for various urban communities. This proposed model efficiently estimated the dengue and assisted in dengue outbreak and control it at its early stages before the outbreak spreads.
3. Novel classifier particularity for the dengue analysis [3], focus to develop a prognostic desire model using the seven counsel signs included by WHO revised Dengue fever portrayal 2009 to choose extraordinary dengue in kids. Systems. In this imminent explanatory examination drove in a tertiary thought network, ceaseless assessing of all children developed 1mo to 12y surrendered with serologically confirmed Dengue was done from May 2015 through August 2016. Subsequent to accepting 27 patients with co-pollutions and co-morbidities, 359 patients were followed up step by step to review clinical and inquire about focus development till discharge/downfall. Free indicators were stomach agony or delicacy, tireless spewing, torpidity, mucosal drain, clinical liquid amassing, hepatomegaly >2 cm and rising hematocrit simultaneous with platelet check. To propose real time dengue risk prediction. It used prediction model instead of traditional statistical model for early warnings.
4. Dengue prediction system using data mining [4]. This paper targets performing Named Entity Recognition to extricate scatter refers to, time articulations and other important highlights from clinical information. They fabricate a model to anticipate the nearness or non-appearance of the dengue illness and performed recurrence examination which relates the event of dengue and the indication of its side effects throughout the months. A lot of commented on release synopses are utilized as a contribution to the proposed framework. Execution measurements considered right now Accuracy, Kappa insights, Mean Absolute Error, Root Mean Square Error, and Relative Absolute Error. It is reasoned that the presentation of the SMO calculations is superior to other algorithms.
5. Dengue fever classification using gene expression data [5]. Dengue fever discovery and arrangement have an essential job because of the ongoing episodes of various types of dengue fever. As of late, the headway in the microarray innovation can be utilized for such an order process. A few examinations have set up that the quality choice stage plays a critical job in the classifier execution. In this manner, the present examination concentrated on recognizing two distinct varieties, specifically, dengue fever (DF) and dengue hemorrhagic fever (DHF). An adjusted bag of-highlights strategy has been proposed to choose the most encouraging qualities in the arrangement procedure.
6. Mining Methods & Techniques for Dengue Detection [6]. Data Mining is centered on getting information from numerous information positions utilizing insightful examination methods. Information mining is a field having different strategies that change over the crude information into valuable data in different research fields. There are two essential objectives for information mining expectations and portrayal. The forecast includes fields or factors in the informational indexes to anticipate obscure or future estimations of different sicknesses prospects. Then again depiction includes finding of example portraying the information that can be available in information base accommodated malady forecast. We can anticipate ailments like hepatitis, Lung malignant growth liver issue, bosom malignant growth or heart infections, diabetes and so on, it helps in finding the examples to choose future patterns in the clinical field. We can use it in finding the information and expectation, identification of Dengue also.
7. Early Detection of Dengue utilizing Machine Learning Techniques [7]. Dengue is a pandemic sickness found in a couple made also as making nations like India. Late research assesses that there were 390 million dengue ailments reliably and predicts that dengue transmission is available all through the tropics, with the most significant risk in the American and the Asian districts. This illness is accomplished by impersonating a mosquito called Aedes. Dengue has become a general issue and it is crucial in all the nations. Early censures of dengue scene will help with diminishing the defilement bother and to control it. Information tunneling is the procedure for the solicitation for ailments like dengue. Right now, the Weka tool compartment is utilized to assess and look at the results.
8. Dengue fever prediction using classification techniques [8]. Data mining is the way toward finding designs in enormous datasets including techniques at the convergence of AI, insights, and frameworks. Information mining is the investigation phase of the "information disclosure in databases" procedure or KDD. It is dangerous for clinical advisors to foresee the dengue chance assault as it is a mind-boggling task that requires understanding and information. The wellbeing segment today contains hidden data that can be significant in deciding. Dengue fever dataset is utilized in arrangement methods to evaluate and look at their presentation. For appropriately ordering our dataset, the proposed technique is contrasted and the choice tree.
9. Dengue Possibility Forecasting Model [9]. Dengue is a quick-rising pandemic-inclined viral ailment, with more than 33% of the total populace in question. To check this issue, it's important to make a prescient framework which can limit the harm and misfortune ahead of time. Our dengue occurrence forecast model fuses Gradient Boosting Regression (GBR) calculation and Mean Square Error (MSE) to measure the exhibition of the model. Dataset for Dengue gives data about the patients enduring with the dengue illness. The Dataset comprises a trait like a temperature, precipitation and so.
10. Dengue Disease Prediction Using SVM and Decision Tree [10] Dataset was gathered with Weka and Net Beans IDE. A choice tree is made utilizing Fisher Filtering and expectation is finished with the assistance of Support Vector Machine (SVM).

**CHAPTER 3**

**SYSTEM ANALYSIS**

**3.1 REQUIREMENT ANALYSIS**

**3.1.1 FUNCTIONAL REQUIREMENTS**

Functional Requirements includes the defined component in our system and describes what features the system must comprise of. Various predictive models for the given problem statement. According to the various marketing, prediction needs described following are functional requirements. Functional Requirements are characterized by the segment of our framework and the features contained in it. Various predictive models are described below:

**Medical Requirements:** In our project, our motive is to help the medical practitioner to detect the dengue in the early stage. It will help to cure the disease very soon. In dengue, most of the patients need external platelets, WBC, etc to make the patients healthy. To check how much our product is successful we have to check the accuracy, precision of our product that we will be going to check in our project.

**Economic requirement**: After streamlining our business requirements, we can influence the business and national economy on a large scale. Employment opportunities are created and a large part of the stock market gets influenced, eventually even affecting the world economy.

**Feature Engineering:** Feature engineering is used here to help in finding out and extracting the new features from the existing ones. It also includes the inclusion of additional features derived from existing ones. Thereby helps in the isolation of key specific information.

**Data Standardization**: To perform zero mean centering accurately data standardization is very important.

**Baseline Modeling**: In the baseline model for each test data point there is a prediction of the train labels as the output and helps perform a general hypothesis testing of complex models. Our baseline model is that of the decision tree.

**3.1.2 NON-FUNCTIONAL REQUIREMENTS**

This is a requirement explaining what system does by making use of specific criteria to evaluate the operations of our system in the best possible manner. They don't describe the system's function but how the system does those functions. The best possible cross-validation score is feasible time complexity. Following are the non-functionality requirements:

**Scalability**: The system should be scalable and not suffer any performance issues while working on any other systems. Scalability help in the handling of the increasing amount with the addition of resource to our system. The system should be more adaptable to the changing needs of the environment as well as to satisfy the user-specific requirements.

**Maintainability:** The maintenance which is required for the system must be low and not cost conducive.

**Reliability:** The system must be a reliable platform while being an effective tool in being efficient and accurate cost prediction for the customer. It describes the dependability as well as the power of our system or a specified component within the system for functioning accurately under any specific conditions for the given period. While satisfying all these conditions the system is said to be reliable.

**Performance:** Efficient performance by the system in less time while maintaining the accuracy of the solution to the problem statement. It encompasses the various techniques that are applied in this system for ensuring that this non-functionality requirement such kind of throughput, memory use, etc. are all satisfied.

**Recoverability:** It is required from a system to able to recover from minor failures and be self-sufficient enough to solve them. The time taken by a system to restart depends upon the specified re-start point and the time taken to recover from them immediately with effect.

**Capacity**: The systems or any component within this system must able to working under immense pressure and should be able to take high loads. Capacity thereby increases the reliability due to its efficiency in handling load and being more scalable in its approach.

**3.2.3 HARDWARE REQUIREMENTS**

Following are the hardware requirements:

1. System: Pentium IV 2.4 GHz.
2. Hard Disk: 45 GB.
3. Floppy Drive: 1.42 MB
4. Monitor: 15 VGA Colour.
5. Mouse: Logitech
6. Ram: 1GB

**3.2.4 SOFTWARE REQUIREMENTS**

1. Operating system: Windows XP/7/10
2. Programming Tool: MATLAB

**CHAPTER 4**

**SYSTEM DESIGN**

**4.1** **ARCHITECTURE AND PROCESS FLOW**

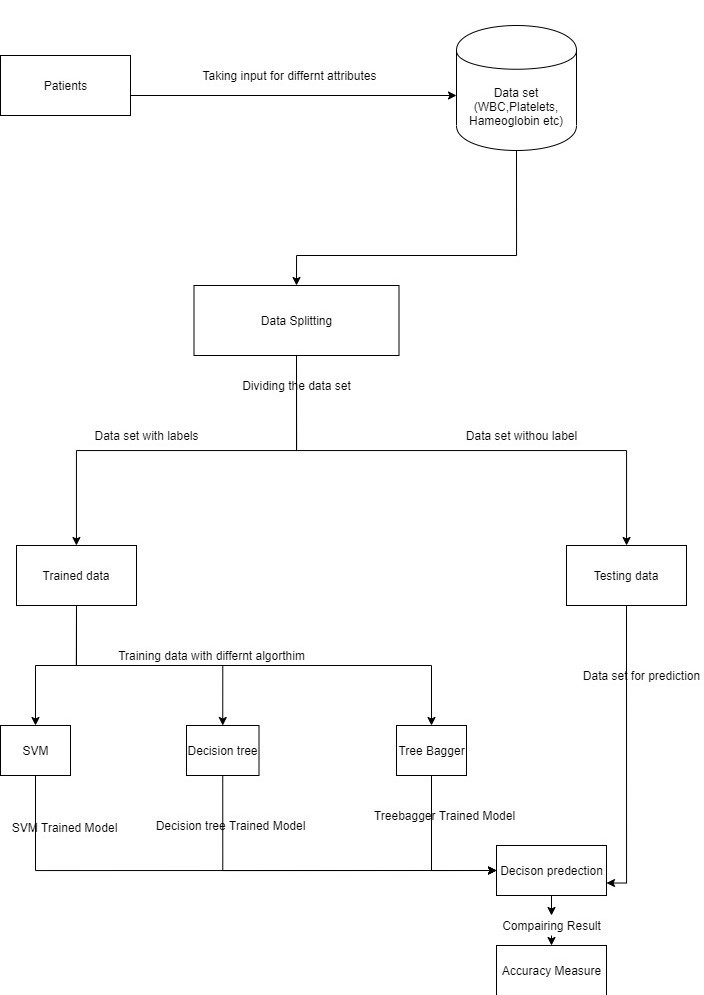


Fig. 4.1 System Architecture

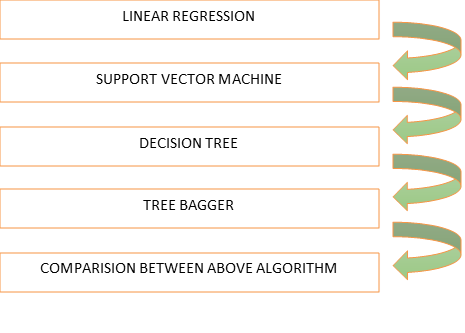


Fig. 4.2 Selection of algorithms

The DFD (information stream graph) additionally, regularly known as the Bubble Chart. It is a clear graphical formalism that can be used to address a system to the extent data to the structure, diverse getting ready finished on this data, and the came about data is come about by this system.

**Data exploration and visualization:**

Beneﬁts of mental image is that it permits the U.S.A. visual access to large amounts of information in simply digested visuals.

Building models:

1. Linear regression
2. Support Vector Machine
3. Decision Tree
4. Tree Bagger
5. Comparison between different algorithms

In the final stage of our model, it will predict whether the patient is having normal or severe dengue.

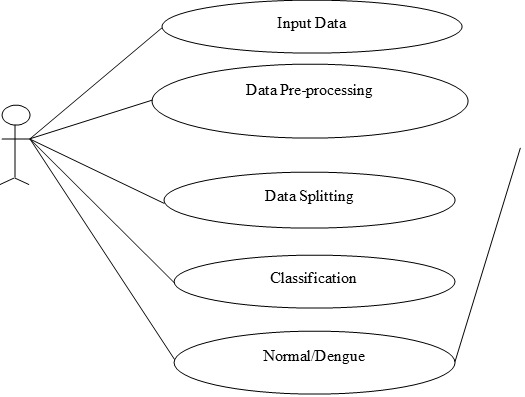


Fig. 4.3 Use Case Diagram

A use case diagram in the Unified Modeling Language (UML) is a behavioral diagram produced after an in-depth use-case analysis. It spurs to present a graphical outline of the convenience gave by a structure with respect to performers, their goals (addressed as use cases), and any conditions between those use cases. The major inspiration driving a use case diagram is to show what structure limits are performed for which performers. The employments of the on-screen characters in the structure can be outlined.

**GOALS**

1. Give customers a prepared to-use, expressive visual showing Language with the goal that they can make and exchange significant models.

2. Give extensibility and specialization parts to grow the middle thoughts.

3. Be self-governing of explicit programming dialects and procedure of improvement.

4. Encourage the development of the OO apparatuses advertise.

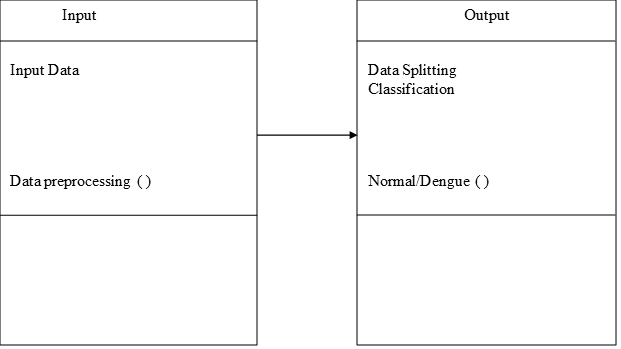


Fig 4.4 Class Diagram

In programming building, a class chart in the Unified Modeling Language (UML) is a kind of static structure graph that portrays the structure of a framework by indicating the framework's classes, their characteristics, activities (or strategies), and the connections among the classes. It clarifies which class contains data.

**4.2 TOOLS DESCRIPTION**

MATLAB is an unmatched language for particular selecting. It encourages calculation, depiction, and programming in a simple to-utilize condition where issues and plans are passed on in like manner numerical documentation.

Standard uses consolidate:

1. Math and count

1. Calculation improvement

1. Information making sure about

1. Displaying, multiplication, and prototyping

1. Information examination, examination, and portrayal

1. Logical and building representations

1. Application improvement, including graphical UI building

MATLAB is an astute structure whose central information portion is a show that doesn't require dimensioning. This licenses you to deal with different explicit getting ready issues, particularly those with framework and vector nuances, in a constrained amount of the time it would take to shape a program in a scalar non-shrewd language, for example, C or Fortran. The name MATLAB addresses a framework investigate office. MATLAB was from the start made to give essential access to orchestrate programming made by the LINPACK and EISPACK experiences. Today, MATLAB motors join the LAPACK and BLAS libraries, installing the bleeding edge in programming for cross segment estimation.

**MATLAB MATHEMATICAL FUNCTIONAL LIBRARY**

This is a huge grouping of computational estimations going from simple limits, like aggregate, sine, cosine, and complex calculating, to logically present day limits like system chat, framework eigenvalues, Bessel limits, and snappy Fourier changes.

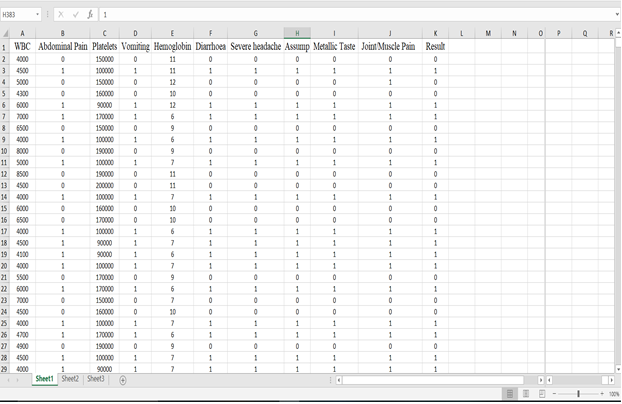
**CHAPTER 5**

**PROPOSED METHODOLOGY**

**MODULE DESCRIPTION**

**1. Dataset Preparation**

The data set is collected from hospitals. We use 10 medical data attributes i.e. WBC, Hemoglobin, Platelets, Abdominal Pain, Vomiting, Diarrhea, Severe headache, Dengue, Metallic Taste, Joint/Muscle Pain for dengue disease prediction.

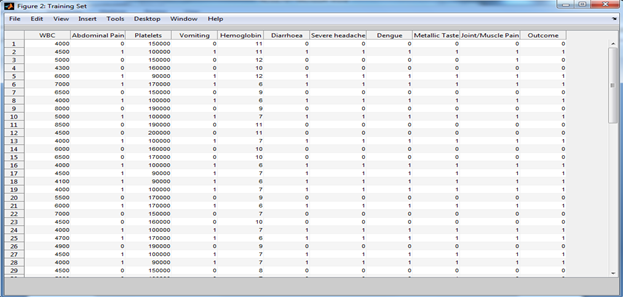
  
Fig. 5.1 Understanding the data set

**2. Information Preprocessing**

The dataset contains unique characteristics and NaN esteems. In programming, one can't process NaN esteems so those qualities are changed into other worth for example numerical worth. NaN esteems are supplanted by mean estimation of segments**.**

**3. Information Splitting**

The parting step is utilized for making the preparation and testing information to the examining procedure. In that, our entire dataset is isolate into preparing and testing information; utilize 80% of information for preparing and 20% of information for testing. Out of a total of 500, the training dataset consists of 396, while the testing dataset consists of 104.

Fig. 5.2 - Training data set

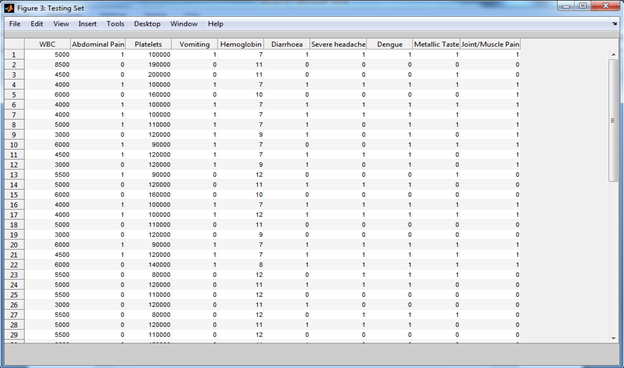
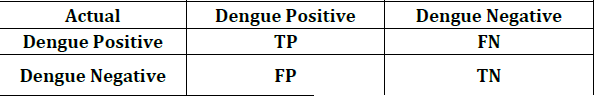


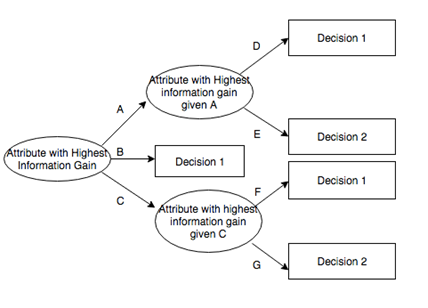
Fig- 5.3 Testing set

**4.Classification**  
In characterization, split preparing and testing information are valuated dependent on AI models. In the first place, preparing information was prepared by utilizing three diverse AI models, for example, Support Vector Machine, Decision Tree and Tree-Bagger. After that testing information are approved dependent on prepared information with a high arrangement precision rate.

To compute Accuracy, precision, and recall, the confusion matrix is needed:  
  
Fig. 5.4 Confusion Matrix

**4.1Support-Vector-Machine**  
A Support Vector Machine is a regulated learning calculation. SVM models the information into k classifications, performing classification and shaping an N-dimensional hyperplane. These models are fundamentally the same as a neural network. Consider a dataset of N measurements. The SVM plots the training information into an N dimensioned space. The training information points are then isolated into k various locales relying upon their marks by hyper-planes of n various measurements. After the testing stage is done, the test focuses are plotted in a comparative N dimensioned plane. Dependent upon the territory the focuses are arranged in.

**4.2Decision-Tree**  
The choice tree is a kind of directed learning calculation that is generally used in arrangement issues. It works for both all out and ceaseless information and yield factors. In our venture, we split the information into in any event two homogeneous sets (or sub-masses) as a result of the most basic splitter/differentiator in input factors. In Decision-Tree inward hub speaks to the test on the trait, the branch depicts the outcome and leaf tells the decision made in the wake of figuring characteristic.

  
Fig. 5.5 Working of Decision Tree

**4.3Tree-Bagger**  
Tree Bagger bags an outfit of decision trees for either grouping or relapse. Bagging represents bootstrap aggregation. Each tree in the troupe is developed on an autonomously drawn bootstrap reproduction of input data. Perceptions excluded from this copy are "out of sack" for this tree. Tree Bagger depends on the Classification Tree and Regression Tree usefulness for developing individual trees. Specifically, Classification Tree and Regression Tree acknowledges the number of highlights selected at random for each decision split as an optional input argument.

**CHAPTER 6**

**SYSTEM TESTING**

## UNIT TESTING

It is the way toward testing every single module created by the designers. The whole program is divided into numerous bundles which comprise of little units of code. It improves the general structure of the module and refactors the code wherever essential. These modules are tried autonomously independent of different modules. They are tried in a successive request also, it checks for repetition. If there should arise an occurrence of redundancy it erases the copy records. It too checks for run time blunder and checks if the connection gave take them to the individual page. Preferred standpoint of performing unit testing is its capacity to check every module exclusively which is supportive in finding the littlest of littlest mistakes. Since unit testing is done at an in all respects early stage the expense of testing is negligible when contrasted with other testing. Modules which are as well enormous for unit testing can be assessed utilizing integration testing.

## INTEGRATION TESTING

This is subsequent stage after unit testing is performed. Once, every module tried autonomously is clear of mistakes, these individual modules are consolidated together and tried in general. The fundamental explanation behind playing out this test is to check for issues when every one of the units are joined. There are diverse manners by which these units can be coordinated. They are:

1. Top Down Integration - Top-down mix joins and tests every one of the modules start to finish. However, one inconvenience of this testing is that it needs more stubs.
2. Bottom Up Integration - The base up methodology is the other way around of top- down approach. Significant modules are tried last which can make issues amid combination.
3. Big-Bang Integration - In this type of testing every one of the functionalities are incorporated and tried at the same time. This methodology is subject to the quantity of modules present. Lesser The modules progressively successful it is.
4. Hybrid Integration – It is a mix of all the above methodologies.

## SYSTEM TESTING

System Testing is the subsequent stage after coordination testing. In this procedure the entire item is tried for issues and mistakes. They are of two kinds:

1. Black box testing
2. White box testing

A case for this is assembling of ballpoint pen. The top, the ink cartridge, the body, the tail is created independently and tried independently (unit testing). Whenever at least two modules are prepared, they are consolidated and Integration Testing is finished. At the point when the total pen is collected, System Testing is finished. It thinks about the entire system as single element.

### Black Box Testing

It is a testing method which is completed by the analyzers. This product can be tried without knowing the inward structure of program. Programming Knowledge isn't expected to do this type of testing procedure. Its fundamental desire is to check for the activity that is performed by the system. It is less tedious. Black box testing is generally called functional test or external testing. It isn't best for algorithm testing. It very well may be tried on preeminent dimensions of testing like acceptance testing.

### White Box Testing

### It is a testing technique which is done by s/w engineers. The usefulness of the program must be known to the designer. Programming learning is an unquestionable requirement to perform White Box Testing. It is generally called inside testing or basic testing. Its principle point is to check program code, circles, conditions, branches and how framework is performing. It tends to be tried on more elevated amounts of testing like acknowledgment testing and acknowledgment testing.

## REGRESSION TESTING

This is a standout amongst the most significant sort of testing with regards to the correct advancement of a product. We can likewise consider it as one significant advance in the Software Development Life Cycle (SDLC). Each product has a particular sort of functionalities which should be refreshed without fail. This is typically done to guarantee its security in all stages. Along these lines, for this to be guaranteed, these functionalities need to refreshed with new bit of code without fail. In this manner, so as to guarantee that the new code doesn't influence the new usefulness, relapse testing is completed. This is normally done by specialists or programming engineers who have profound comprehension of the product activities in and out.

## SMOKE TESTING

It is additionally one angle to ensure that the usefulness is simply working fine independent of the new code that is added to change it. A standout amongst the most significant motivation to play out this type of testing is to expel every one of those lines of code that isn't required any longer and make sure that they try not to influence the usefulness of the product. It covers the greater part of the critical elements of the programming however does not dissect them in detail. The outcome of this test is used to pick whether to proceed with further testing. If the smoke test passes, continue with further testing. In case it misses the mark, end further tests and demand another structure with the required fixes.

## ACCEPTANCE-TESTING This is the last period of testing which is performed by or before customers. This testing is fundamentally done to check whether the created item fulfills the customer's necessity. They are 4 distinctive manners by which acknowledgment testing can be performed. They are Client acceptance testing, Business acceptance testing, Alpha testing and Beta testing.

Table 6.1: Different Testcases with their obtained outcomes

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TEST CASE ID | TEST CASE DESCRIPTION | STEP DETAILS | EXPECTED RESULT | ACTUAL RESULT | STATUS |
| 001 | Check if the data is being correctly mapped to the dataframe | We check if the dataframe contains data or not | The assertion is to be True as the dataframe does contains the imported  data | The data is present in the dataframe. | Pass |
| 002 | Check if the number of rows and columns in the data frame are matching with the entries in the dataset | Check if there are 500  rows and 10 columns in the data frame | The assertion is to be True as the data frame must contain 10 columns  and 500 rows | The dataframe consists of 10  columns and 500 rows. | Pass |
| 003 | Check if there are any null values/empty spaces in the table | Check if there are any null values/empty spaces in the table | The assertion must return False as there aren’t any empty spaces in the  data frame | The assertion returns False | Pass |
| 004 | Check if the total number of cases are the correct number | Check if total=number of positive + number of negative | The assertion must return True, as the total is 500 | The assertion returned is True, proving the total to be  500 | Pass |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 005 | Check if the accuracy rate  for training is more than 70% | Check if the accuracy rate  for SVM training is more than 70% | The assertion  must return True, as the accuracy expected is more than  70% | The assertion  returned is True as the accuracy rate is more than 70% | Pass |
| 006 | Check if the accuracy rate  for testing is more than 70% | Check if the accuracy rate  for SVM testing is more than 70% | The assertion  must return True, as the accuracy expected is  more than 70% | The assertion  returned is True as the accuracy rate is more than 70% | Pass |
| 007 | Check if the accuracy rate  for training is more than 80% | Check if the accuracy rate  for Decision Tree training is more than 80% | The assertion  must return True, as the accuracy expected is more than  80% | The assertion  returned is True as the accuracy rate is more than 80% | Pass |
| 008 | Check if the accuracy rate for testing is more than 80% | Check if the accuracy rate for Decision Tree testing is more than 80% | The assertion must return True, as the accuracy expected is more than  80% | The assertion returned is True as the accuracy rate is more than 80% | Pass |
| 009 | Check if the accuracy rate for testing is more than 90% | Check if the accuracy rate for Tree Bagger testing is more than 90% | The assertion must return True, as the accuracy expected is more than  90% | The assertion returned is True as the accuracy rate is more than 90% | Pass |
| 010 | Check if the accuracy rate for training is more than 90% | Check if the accuracy rate for Tree Bagger training is more than 90% | The assertion must return True, as the accuracy expected is more than  90% | The assertion returned is True as the accuracy rate is more than 90% | Pass |

**Training and Testing:**

We can isolate the information into a preparation informational collection and testing informational collection. Most informational collection could be utilized for preparing segment of the information is utilized for testing. The preparation information is utilized to process and plan the model according to necessity. After a model has been finished utilizing coming down set, you test the model by making expectations against the test set.

For testing of the preparation informational collection we have utilized the cross-approval procedure. Cross approval strategy is more qualified for this kind of prescient framework. In cross approval strategy the informational collection is divided with the goal that the model is prepared alright. In cross approval model the subsets are prepared for progressively precise outcomes. Dividing and sub setting are the central point which are to be thought of while performing cross approval. There are two kinds of cross approvals initially is comprehensive cross approval second one is non thorough cross approval.

We have used cross validation in random forest with feature selection. It has been depicted in visualization. Training data set consists of 120 entries among 11 variables. Test data set consists of 384 entries among 10 variables. 11th variable in the training data set is kept as the response variable which needs to be predicted in the test data set. This is one such example where cross validation is used for testing. Our testing part is mainly done through cross validation.

**Cross Validation**

Cross validation is the technique which does validation by making partitions. In our project cross validation has helped to overcome the problem of overfitting which is a very common problem for machine learning algorithm. Cross validation is also called as rotational estimation. It is very important step in training phase. It has helped us to maintain the quality of our model. There are two types of cross validation first is exhaustive and the other is non exhaustive.

**Analytics report:**

Generating the end result, and drafting and analyzing it by plotting them diagrammatically. Also we ensure that the model is not over fitted and provides desired accurate result.

**CHAPTER 7**

**EXPLORATORY RESULTS**

Here, we tell the characterized outcome from different forecast model. We utilized various parameters in order for making examinations with various models, the parameters for example Exactness, Precision, and Recall.

|  |  |  |
| --- | --- | --- |
| Genuine/Predicted | Positive | Negative |
| Positive | TP | FN |
| Negative | FP | TN |

Table 7.1: Disarray Network

Based on the above prediction, we evaluate the parameters i.e. accuracy, precision and recall.

1. Exactness: It gauges the examination of TP and TN to the complete no. of test pictures.

(3)

1. Precision: It is the estimation examination of genuine positive to the total estimation of genuine positive and bogus positive rate. It is given in eqn. (4)

(4)

1. Review: It is the estimation examination of genuine positive rate to the total estimation of the genuine positive and bogus negative rate. It is given in eqn. (5).

(5)

**OUTPUT TABLE:**

|  |  |  |  |
| --- | --- | --- | --- |
| Method | SVM | DT | TB |
| Exactness | 0.8923 | 0.8769 | 0.9230 |
| Precision | 0.9310 | 0.8529 | 0.9354 |
| Review | 0.8437 | 0.9062 | 0.9062 |

Table: Quantitative Evaluation with different models

## 7.1 COMPARISON GRAPHS

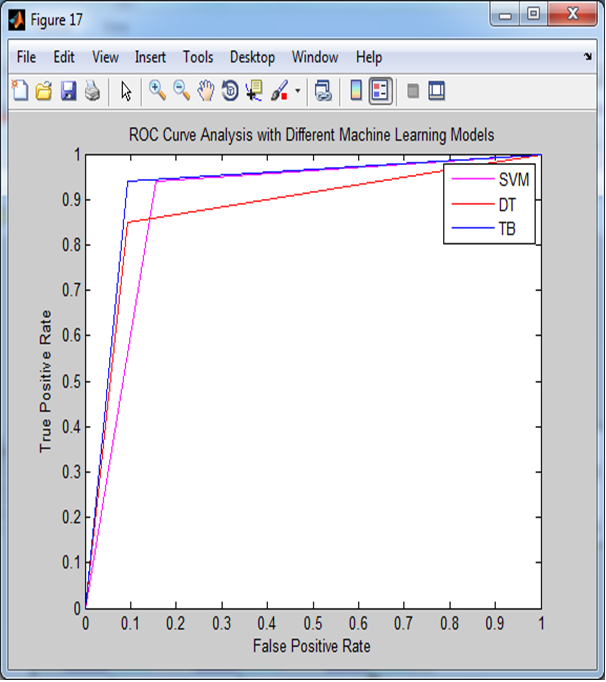


Fig. 6.1 ROC Curve Analysis vs Machine Learning Models

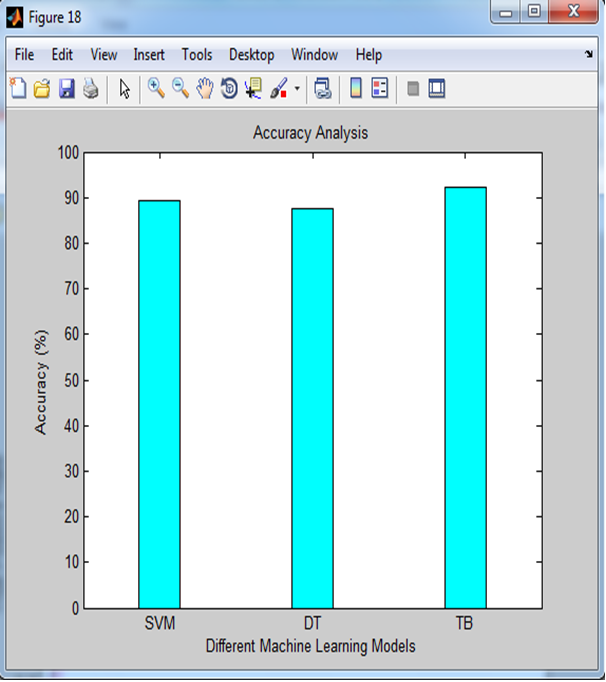


Fig. 6.2 Graphical Analysis of Exactness

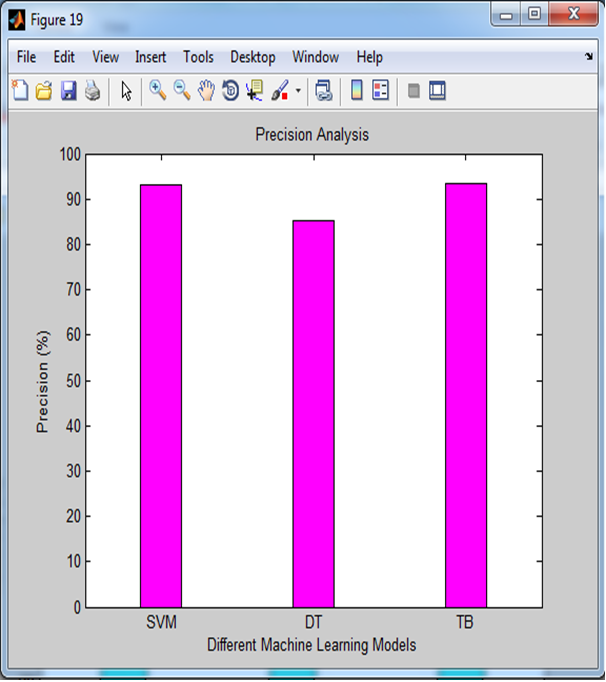


Fig. 6.3 Graphical Analysis of Precision

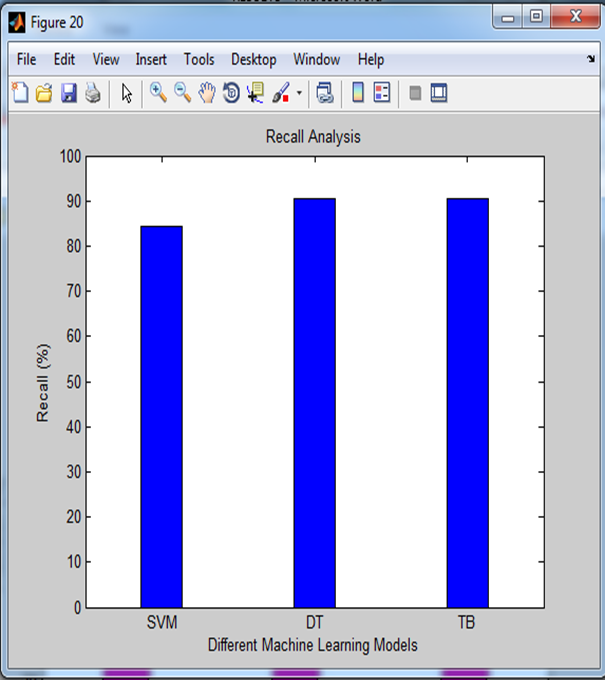


Fig. 6.4 Graphical Analysis of Review

We can infer from the above results that Accuracy, Precision, and Recall of the Support Vector Machine (SVM) algorithm are 0.8923, 0.9310 and 0.8437 respectively. By comparing it with Decision Tree we can infer that Accuracy, Precision, and Recall of the Decision Tree (DT) algorithm are 0.8769, 0.8529, 0.9062 respectively, which is less in comparison with SVM.

At last, by examining that output from Tree Bagger Algorithm (TB) we can infer that

Accuracy, Precision, and Recall of the Tree Bagger are 0.9230, 0.9354 and 0.9062 respectively.

Hence, our best algorithm is **Tree Bagger** which is giving the most accurate results.

**7.2 OUTPUT SNAPSHOTS**

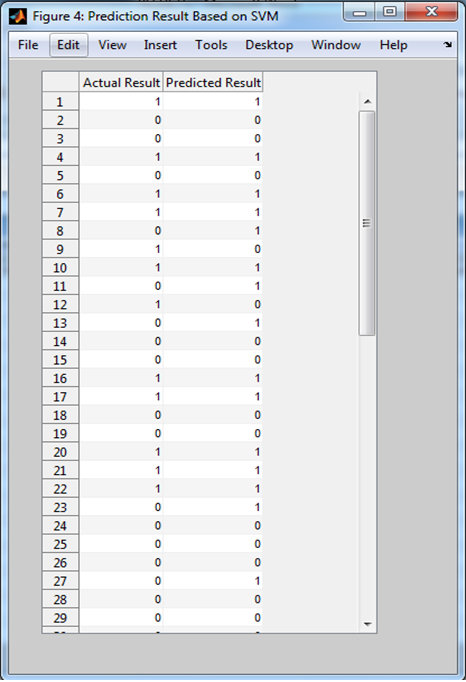


Fig. 6.5 Prediction result based on SVM

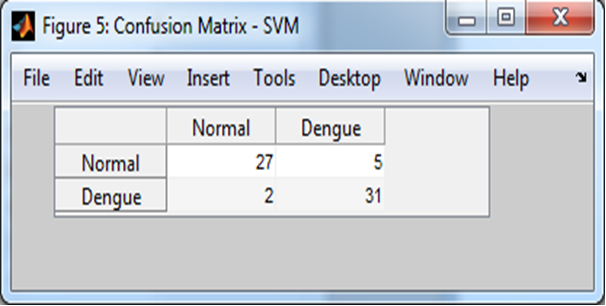


Fig. 6.6 Confusion Matrix of SVM

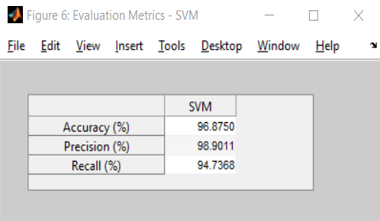


Fig 6.7 Evaluation Metrics of SVM

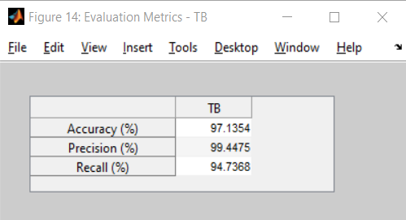


Fig 6.8 Evaluation Metrics of TB

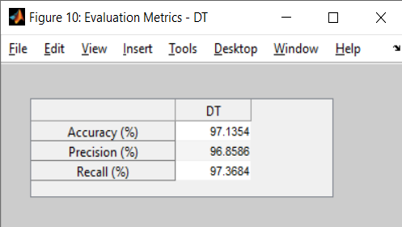


Fig. 6.9 Evaluation Metrics of DT

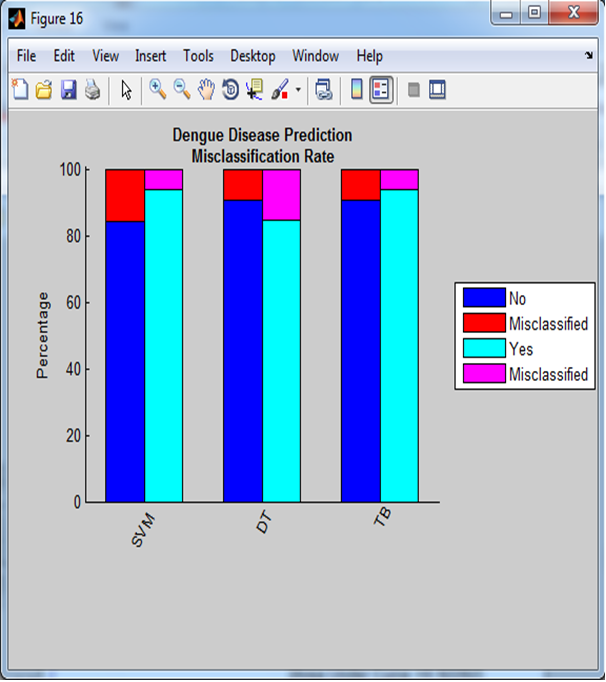


Fig. 6.10 Misclassification curve

**CHAPTER 8**

**CONCLUSION**

The dengue disease prediction is the issue of expectation investigation. In this work we analyzed the different AI calculations for example Support Vector Machine, Decision Tree and Tree Bagger for predicting of dengue disease.

Expectation of dengue infection followed the progression as information arrangement in that information was gathered from emergency clinics.

Information preprocessing for supplanting of NaN esteem. Information preprocessing is a necessary advance in Machine Learning as the nature of information and the valuable data that can be gotten from it legitimately influences the capacity of our model to learn; thusly, it is critical that we preprocess our information before taking care of it into our model.

Information parting is that in which entire dataset split into preparing and testing. Data is isolated into preparing and testing set. For preparing 80% of the information is utilized and for testing 20% information is feed into the machine.

At long last various models based preparing information are prepared and testing information are approved by the prepared information in characterization step. Finally, various AI models are thought about in terms of finding of quantitative assessment measurements, for example, precision, accuracy and recall.

The simplest of all the methods is decision tree. It tends to be utilized for both order and relapse issues and provide us a useful and simple tool for interpretation. The major contrast among classification and regression trees is the information type of the objective variable. At the point when our objective variable is a discrete set of values, we have a classification tree.

On the other hand, a regression tree has its objective variable to be continuous values. In the arrangement issues, our leaves are class names, while in the relapse issues, our leaves are bins of continuous values of the objective variable. All perceptions falling into a similar receptacle have the equivalent anticipated worth.

Presently for Support Vector Machine (SVM) works modestly well when there is clear edge from of partition between classes. SVM is progressively effective in high dimensional spaces.

SVM is viable in circumstances where number of estimations is more than the amount of tests.

SVM is sensibly memory practical. Regardless, SVM tallies isn't reasonable for huge datasets. SVM doesn't perform well indeed, when the dataset has more uproar for instance target classes are covering. In conditions where number of features for every datum point beats the measure of preparing information test, the SVM will disregard to meet wants. Finally, Tree bagger is the calculation which is the most impressive and gives the best yield as far as exactness, accuracy and review. Packing is the sort of system which is utilized when the objective is to lessen the change of the choice tree. Our goal here is that to frame a few subsets of information from preparing test picked haphazardly with substitution and afterward every assortment of the subset information is utilized to prepare their choice tree. After comparing all these we observed that Tree Bagger is by far giving us the best prediction in terms of accuracy, precision and recall, and hence it is the best algorithm among all the three algorithms.

**CHAPTER 9**

**FUTURE ENHANCEMENTS**

The proposed work can be investigated in terms of practicality and exactness, so recognizable proof and forecast of dengue play an essential role in saving an individual's life since heaps of personal life are in danger and bunches of individuals died because of dengue since dengue isn't analyzed or individual having less equipment's.

Our work right currently is to make an authority system that uses AI calculations like Support Vector Machine (SVM), Decision Tree and Tree Bagger (TB). With the help of these methodology, we are prepared for making an ace system that urges us to predict dengue diseases dependent on a tainted individual's side effects.

Forecast or Prediction of dengue is a life-saving project and with the assistance of this, we can save many people's lives.

There are loads of work pending in this field right now and many specialists are looking for new techniques in this work to make this prediction more and more accurate.

Our future aim is to test this model on larger datasets to work upon its abnormalities and improve the project so that it will get better in terms of precision and accuracy. We hope it will be useful in saving millions of lives.

**CHAPTER 10**

**REFERENCES**

[1] Bhatt et al., “The global distribution and burden of dengue,” Nature vol. 496, no. 7446, p. 504, 2013.

[2] Ta-Chien Chan, Tsuey-Hwa Hu, Jing-Shiang Hwang, "Daily forecast of dengue fever incidents for urban villages in a city", International Journal of Health Geographics,2015.

[3] Tuan et al., “Sensitivity and specificity of a novel classifier for the early diagnosis of dengue,” PLoS neglected tropical diseases, vol. 9, no. 4, p. e0003638, 2015.

[4] Nandini V, Sriranjitha R, Yazhini TP, "Dengue detection and prediction system using data mining with frequency analysis", Computer Science & Information Technology (CS & IT), 2016.

[5] Chatterjee et al., “Clinical application of modified bag-of-features coupled with hybrid neural-based classifier in dengue fever classification using gene expression data,” Medical & biological engineering & computing, pp. 1–12, 2017.

[6] Dave Kaveri Atulbhai and Shilpa Serasiya, “A Survey: Prediction & Detection of Dengue – Mining Methods & Techniques, IJARIIE–ISSN (O)-2395-4396, Vol-3, Issue-2, 2O17.

[7] Early Detection of Dengue Using Machine Learning Algorithms 1N.Rajathi, 2S.Kanagaraj, 3R.Brahmanambika and 4K.Manjubarkavi, 1,2,3,4Kumaraguru College of Technology, Coimbatore-49.   
  
[8] Dengue fever prediction using classification techniques, R. Sanjudevi1, D. Savitha2

[9] Dengue Possibility Forecasting Model using Machine Learning Algorithms P.Muhilthini\*1, B.S. Meenakshi\*2, S.L. Lekha\*3, S.T. Santhanalakshmi\*4.

[10] Dengue Disease Prediction Using Decision Tree and Support Vector Machine, Dr.Arun Kumar.P.M, Chitra Devi.B, Kar-thick.P, Ganesan.M and Madhan.A.S

**APPENDIX- I**

**Codes for Data Preparation**

**1. Loading Data:**

A = xlsread('dataset.xlsx');

f = figure('name', 'Dengue Disease Prediction Dataset', 'Position', [200 100 1000 600]);

cnames ={'WBC', 'Abdominal Pain', 'Platelets', 'Vomiting', 'Hemoglobin', 'Diarrhea', 'Severe headache', 'Dengue', 'Metallic Taste', 'Joint/Muscle Pain', 'Outcome'};

t = uitable('Parent', f, 'Data', A, 'ColumnName', cnames, 'Position', [0 40 1000 550]);

[nrows, ncols] = size(A);

category = true(1, ncols);

catPred = category(1: end-1);

**2. Dataset Splitting:**

% Training set

Xtrain = A(1:120,1:10);

Ytrain = A(1:120,11);

f = figure('name', 'Training Set', 'Position',[200 100 1000 600]);

cnames = {'WBC', 'Abdominal Pain', 'Platelets', 'Vomiting', 'Hemoglobin', 'Diarrhea', 'Severe headache', 'Dengue', 'Metallic Taste', 'Joint/Muscle Pain', 'Outcome'};

t = uitable('Parent', f, 'Data', [Xtrain Ytrain], 'ColumnName', cnames,   
‘Position', [0 40 1000 550]);

% Test set

Xtest =A(121: end, 1:10);

Ytest = A(121: end, 11);

f = figure('name', 'Testing Set', 'Position', [200 100 1000 600]);

cnames = {'WBC', 'Abdominal Pain', 'Platelets', 'Vomiting', 'Hemoglobin', 'Diarrhea', 'Severe headache', 'Dengue', 'Metallic Taste', 'Joint/Muscle Pain'};

t = uitable('Parent', f, 'Data', Xtest, 'ColumnName', cnames, 'Position', [0 40 1000 550]);

**APPENDIX- II**

**1. Code for Classification of Data**

In classification, split trained and tested data are valuated dependent on machine learning models. To start with, training data is trained by utilizing three diverse machine learning models, i.e. Support Vector Machine, Decision Tree and Tree Bagger. After that testing data is approved dependent on training data with high accuracy rate.

**1.1 Support Vector Machine (SVM)**% Train the classifier  
svmStruct = svmtrain(Xtrain, Ytrain, 'showplot', true);  
  
% Make a prediction for the test set  
C = svmclassify(svmStruct, Xtest, 'showplot', true);  
  
% Compute the confusion matrix  
disp('For SVM');  
C\_SVM = confusionmat(Ytest, C)  
C\_SVM1 = bsxfun(@rdivide, C\_SVM, sum(C\_SVM,2)) \* 100;  
f = figure('name', 'Prediction Result Based on SVM', 'Position', [500 100 400 600]);  
cnames = {'Actual Result', 'Predicted Result'};  
t = uitable('Parent', f, 'Data', [Ytest C], 'ColumnName', cnames,…  
 'Position', [30 40 300 550]);  
f = figure('name', 'Confusion Matrix - SVM', 'Position',[500 400 400 100]);  
cnames = {'Normal', 'Dengue'};  
rnames = {'Normal', 'Dengue'};  
t = uitable('Parent', f, 'Data', C\_SVM, 'ColumnName', cnames, 'RowName', rnames,... 'Position', [30 40 300 60]);

**1.2 Decision Tree**

% Train the classifier

t = ClassificationTree.fit(Xtrain,Ytrain,'CategoricalPredictors',catPred);

% Make a prediction for the test set

Y\_t = t.predict(Xtest);

% Compute the confusion matrix

disp('For DT');

C\_DT = confusionmat(Ytest,Y\_t)

C\_DT1 = bsxfun(@rdivide,C\_DT,sum(C\_DT,2)) \* 100;

f = figure('name', 'Prediction Result Based on DT', 'Position',[500 100 400 600]);

cnames = {'Actual Result', 'Predicted Result'};

t = uitable('Parent', f, 'Data', [Ytest Y\_t], 'ColumnName', cnames,...

'Position', [30 40 300 550]);

f = figure('name', 'Confusion Matrix - DT', 'Position',[500 400 400 100]);

cnames = {'Normal', 'Dengue'};

rnames={'Normal', 'Dengue'};

t = uitable('Parent', f, 'Data', C\_DT, 'ColumnName', cnames, 'RowName', rnames,...

'Position', [30 40 300 60]);

**1.3 Tree Bagger**

cost = [0 1 5 0];

opts = statset('UseParallel', true);

% Train the classifier

tb = TreeBagger(150, Xtrain, Ytrain, 'method', 'classification', 'categorical', catPred, 'Options', opts, 'OOBVarImp', 'on', 'cost', cost);

% Make a prediction for the test set

Y\_tb = tb.predict(Xtest);

Y\_tb = str2double(Y\_tb);

% Compute the confusion matrix

disp('For TB');

C\_TB = confusionmat(Ytest,Y\_tb)

C\_TB1 = bsxfun(@rdivide, C\_TB, sum(C\_TB,2)) \* 100;

f = figure('name', 'Prediction Result Based on TB', 'Position',[500 100 400 600]);

cnames = {'Actual Result', 'Predicted Result'};

t = uitable('Parent', f, 'Data', [Ytest Y\_tb], 'ColumnName', cnames,...

'Position', [30 40 300 550]);

f = figure('name', 'Confusion Matrix - TB', 'Position', [500 400 400 100]);

cnames = {'Normal', 'Dengue'};

rnames={'Normal', 'Dengue'};

t = uitable('Parent', f, 'Data', C\_TB, 'ColumnName', cnames, 'RowName', rnames,...

'Position', [30 40 300 60]);

**2. Code for Accuracy, Precision and Recall**

A11 = length(Ytest);

**2.1 SVM**  
TP = C\_SVM(1,1);  
FN = C\_SVM(1,2);  
FP = C\_SVM(2,1);  
TN = C\_SVM(2,2);  
disp('For SVM');  
Accuracy1 = ((TP+TN) / A11). \* 100  
Precision1 = (TP / (TP+FP)). \* 100  
Recall1 = (TP / (TP+FN)). \* 100  
f = figure('name', 'Evaluation Metrics - SVM', 'Position', [500 400 400 150]);  
cnames = {'SVM'};  
rnames = {'Accuracy (%)', 'Precision (%)', 'Recall (%)};

t = uitable('Parent', f, 'Data', [Accuracy1; Precision1; Recall1], 'ColumnName', cnames, 'RowName', rnames,…  
'Position', [30 30 300 90]);

**2.2 DecisionTree**  
TP = C\_DT(1, 1);  
FN = C\_DT(1, 2);  
FP = C\_DT(2, 1);  
TN = C\_DT(2, 2);  
disp('ForDT');  
Accuracy2 = ((TP+TN) / A11). \*100  
Precision2 = (TP/(TP+FP)). \*100  
Recall2 = (TP/(TP+FN)). \*100  
f = figure('name', 'Evaluation Metrics - DT', 'Position', [500 400 400 150]);  
cnames = {'DT'};  
rnames = {'Accuracy (%)', 'Precision (%)', 'Recall (%)'};  
t = uitable('Parent', f, 'Data', [Accuracy2; Precision2; Recall2], 'ColumnName', cnames,… 'RowName', rnames, 'Position', [30 30 300 90]);

**2.3 Tree Bagger**

TP = C\_TB(1, 1);

FN = C\_TB(1, 2);

FP = C\_TB(2, 1);

TN = C\_TB(2, 2);

disp('For TB');

Accuracy3 = ((TP+TN) / A11). \*100

Precision3 = (TP / (TP+FP)). \*100

Recall3 = (TP / (TP+FN)). \*100

f = figure('name', 'Evaluation Metrics - TB', 'Position', [500 400 400 150]);

cnames = {'TB'};

rnames = {'Accuracy (%)', 'Precision (%)', 'Recall (%)'};

t=uitable('Parent', f, 'Data', [Accuracy3; Precision3; Recall3], 'ColumnName', cnames, 'RowName', rnames, 'Position', [30 30 300 90]);

1. **ROC Curve**
   1. **SVM**

[xx1, yy1, ~, auc] = perfcurve(Ytest,C,1);

figure;

plot(xx1, yy1, 'b-',...

'LineWidth',1,...

'MarkerSize',5,...

'MarkerEdgeColor','r',...

'MarkerFaceColor',[1 1 0]);

hold on

xlabel('False positive rate');

ylabel('True positive rate')

title('ROC Curve for SVM - Predicted vs. Actual (Test Set)')

text(0.5,0.25,{'SVM',strcat('Area Under Curve = ',num2str(auc))},'EdgeColor','k');

* 1. **Decision Tree**

[xx2,yy2,~,auc] = perfcurve(Ytest,Y\_t,1);

figure;

plot(xx2,yy2, 'b-',...

'LineWidth',1,...

'MarkerSize',5,...

'MarkerEdgeColor','r',...

'MarkerFaceColor',[1 1 0]);

hold on

xlabel('False positive rate');

ylabel('True positive rate')

title('ROC Curve for DT - Predicted vs. Actual (Test Set)')

text(0.5,0.25,{'DT',strcat('Area Under Curve = ',num2str(auc))},'EdgeColor','k');

* 1. **Tree Bagger**

[xx3,yy3,~,auc] = perfcurve(Ytest,Y\_tb,1);

figure;

plot(xx3,yy3, 'b-',...

'LineWidth',1,...

'MarkerSize',5,...

'MarkerEdgeColor','r',...

'MarkerFaceColor',[1 1 0]);

hold on

xlabel('False positive rate');

ylabel('True positive rate')

title('ROC Curve for TB - Predicted vs. Actual (Test Set)')

text(0.5,0.25,{'TB',strcat('Area Under Curve = ',num2str(auc))},'EdgeColor','k');

1. **Comparison**

Cmat = [ C\_SVM1 C\_DT1 C\_TB1];

labels = {'SVM ', 'DT ','TB'};

comparisonPlot( Cmat, labels )

plot(xx1,yy1, 'm-',...

'LineWidth',1,...

'MarkerSize',5,...

'MarkerEdgeColor','r',...

'MarkerFaceColor',[1 1 0]);

hold on

plot(xx2,yy2, 'r-',...

'LineWidth',1,...

'MarkerSize',5,...

'MarkerEdgeColor','r',...

'MarkerFaceColor',[1 1 0]);

hold on

plot(xx3,yy3, 'b-',...

'LineWidth',1,...

'MarkerSize',5,...

'MarkerEdgeColor','r',...

'MarkerFaceColor',[1 1 0]);

hold on

title('ROC Curve Analysis with Different Machine Learning Models');

xlabel('False Positive Rate')

ylabel('True Positive Rate')

legend('SVM','DT','TB')

figure,

Z=[Accuracy1 Accuracy2 Accuracy3];

str={'SVM','DT','TB'};

bar(Z,0.27,'c');

set(gca, 'XTickLabel',str, 'XTick',1:numel(str));

xlabel('Different Machine Learning Models');

ylabel('Accuracy (%)');

title('Accuracy Analysis');

figure,

Z=[Precision1 Precision2 Precision3];

str={'SVM','DT','TB'};

bar(Z,0.27,'m');

set(gca, 'XTickLabel',str, 'XTick',1:numel(str));

xlabel('Different Machine Learning Models');

ylabel('Precision (%)');

title('Precision Analysis');

figure,

Z=[Recall1 Recall2 Recall3];

str={'SVM','DT','TB'};

bar(Z,0.27,'b');

set(gca, 'XTickLabel',str, 'XTick',1:numel(str));

xlabel('Different Machine Learning Models');

ylabel('Recall (%)');

title('Recall Analysis');

**PAPER PUBLICATION STATUS**

Publication process not yet started

**PLAGIARISM REPORT**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**  (Deemed to be University u/s 3 of UGC Act, 1956) | | | | | |
| **Office of Controller of Examinations** | | | | | |
| REPORT FOR PLAGIARISM CHECK ON THE DISSERTATION **/** PROJECT REPORTS FOR UG**/** PG PROGRAMMES | | | | | |
| 1 | Name of the candidate (**IN BLOCK LETTERS**) | | MOHNISH KHERA | | |
| 2 | Address of the candidate | | 16, SBI Colony, Barkat Nagar, Tonk Phatak, Jaipur-302015 (RAJ) | | |
| 3 | Registration number | | RA1611008010285 | | |
| 4 | Date of Birth | | 22/08/1998 | | |
| 5 | Department | | Information Technology | | |
| 6 | Faculty | | Engineering and Technology | | |
| 7 | Title of the Dissertation **/** Project | | DENGUE DISEASE PREDICTION & ANALYSIS USING MACHINE LEARNING | | |
| 8 | Whether the above dissertation is done by | | ~~Individual~~ **/** Group   1. If group, number of students: 4 2. Name and Register Numbers of other candidates:   1. Vaibhav Dadhich (RA1611008010189)  2. Manu Anand (RA1611008010441)  3. Pranav Jha (RA1611008010677) | | |
| 9 | Name and address of the Supervisor**/** Guide | | Mrs. G. Geetha  **Email ID**: [geethag@srmist.edu.in](mailto:geethag@srmist.edu.in)  **phone:** 91763 89334 | | |
| 10 | Name and address of the C0-Supervisor**/** Co-guide (if any) | | NIL | | |
| 11 | Software used | | Turnitin | | |
| 12 | Date of Verification | | 02/05/2020 | | |
| 13 | **Plagiarism Details:** | | | | |
| **Chapter** | **Title of the Chapter** | **Percentage of similarity index (including self citations)** | | **Percentage of similarity index (excluding self citations)** | **Percentage of plagiarism excluding Quotes, Bibliography, etc** |
| **1** | Introduction | 5 % | | 2 % | 4 % |
| **2** | Literature Review | 1 % | | <1 % | 0 % |
| **3** | System Analysis | 7 % | | 4 % | 7 % |
| **4** | System Design | 0% | | <1 % | 0 % |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **5** | Proposed Methodology | 2 % | 2 % | 2 % |
| **6** | System Testing | 3% | 2 % | 3 % |
| **7** | Results | 6 % | 6 % | 6 % |
| **8** | Conclusion | 0 % | 0 % | 0 % |
| **9** | Future Enhancements | 0 % | 0 % | 0 % |
| **Appendices** | |  |  |  |
| We declare that the above information has been verified and found true to the best of our knowledge. | | | | |

|  |  |
| --- | --- |
| Signature of the candidate | Name and Signature of the Staff who uses the  plagiarism software |
| Name and Signature of Guide:  Mrs. G. Geetha | Name and Signature of Co - Guide |
| Name and Signature of Head of Department  Dr. G. Vadivu |  |