

Dengue Prediction Using Multilayer Perceptron - A Machine Learning Approach

V. Janani¹, N. Maadhuryaa², D. Pavithra³, S. Ramya Sree⁴

¹Associate Professor, Dept. of Computer Science and Engineering, Adhityamaan College of Engg., Hosur, India

^{2,3,4}Student, Dept. of Computer Science and Engineering, Adhityamaan College of Engineering, Hosur, India

Abstract: Machine learning (ML) is that the application of AI (AI) that has systems the power to be told automatically from experience. The first aim of ML is to permit computers learn without human intervention or assistance. It easily identifies trends and patterns, continuous improvement, handling multidimensional and multi-variety data. Dengue is that the most virally occurring mosquito-borne disease in recent days. The Multilayer Perceptron (MLP) algorithm in Machine Learning is employed to realize accuracy in analysing and predicting dengue disease. The parameters of the dengue integrated model are identified using an optimization-based methodology in multiple stages. The prediction of dengue using the MLP algorithm is allotted by three phases. The initial phase of dengue prediction is data visualization and pre-processing which is implemented by SMO (Sequential Minimal Optimization). SMO is an algorithm for solving the quadratic programming problem that arises during the training of Support Vector Machines (SVM). The second phase is that the MLP feature selection algorithm which has a leverage backward logistic regression risk analysis. The ultimate phase includes feature reduction by MLP could be a part of dimensionality reduction within the dataset. The implementation of dengue prediction is completed by the WEKA tool. WEKA tool could be a collection of machine learning algorithms for data processing tasks. This tool will be applied to a dataset directly or called from the java code and contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Keywords: Data processing (DM), Machine Learning (ML), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Sequential Mining Optimization (SMO), WEKA tool.

1. Introduction

A. Data mining concept

Data Mining is an analytic process. It's designed to explore great deal of dataset typically business or market-related in search of consistent patterns. The concept of data mining is becoming increasingly popular as a business information management tool where it's expected to reveal knowledge structures which is able to guide decisions in conditions of limited certainty. Recently, there has been increased interest in developing new analytic techniques. Processing continues to be supported the conceptual principles of statistics including the conventional Exploratory Data Analysis (EDA) and modelling and it shares with them both some components of its general approaches and specific techniques Accordingly the ultimate

word goal of data mining is prediction - and predictive is that the foremost typical variety of processing and one that has the foremost direct business applications. The strategy of DM consists of three stages:

- 1) The initial exploration
- 2) A model building or pattern identification with validation/verification
- 3) Deployment.

2. Literature review

Natural Medical processing has attracted substantial attention in many applications and research areas. Many of the prevailing approaches are supported calculating distances among the points within the dataset. Besides, current datasets usually have an outsized number of dimensions. These datasets tend to be sparse, and traditional concepts like Euclidean distance or nearest neighbor become unsuitable.

Dietary fat reduction and dengue disease outcome: results from the women's intervention nutrition study (WINS)

[1] Blackburn G L has proposed, an algorithm to perform outlier detection on time-series data is developed, the intelligent outlier detection algorithm (IODA). This treats a statistic as an image and segments the image into clusters of interest, like "nominal data" and "failure mode" clusters. The algorithm uses density clustering techniques to identify sequences of coincident clusters in both the time domain and delay space, where the delay space representation of the statistic consists of ordered pairs of consecutive data points taken from the statistic. "Optimal" clusters that contain either mostly nominal or mostly failure-mode data are identified in both the time domain and delay space.

Cancer genetics, (cancer treatment and research). Berlin: springer

[3] Boris Pasche, has proposed Interestingness measures play a really important role in processing, no matter the type of patterns being mined. These measures are intended for selecting and ranking patterns in line with their potential interest to the user. Good measures also allow the time and space costs of the mining process to be reduced. The interestingness measures for rules and summaries, classifies them from several perspectives, compares their properties, identifies their roles within the information processing process, gives strategies for selecting

appropriate measures for applications, and identifies opportunities for future research during this area.

Predicting dengue disease survivability: a comparison of three processing methods

[5] Delen D has proposed it proves an bound for the memory consumption which allows the invention of all outliers by scanning the dataset 3 times. The bound seems to be extremely low in practice. Since the actual memory capacity of a practical DBMS is commonly larger, we develop a novel algorithm, which integrates our theoretical findings with carefully designed heuristics that leverage the additional memory to spice up I/O efficiency.

3. Proposed system

The proposed framework SMO based on disease prediction is shown to be effective in addressing this prediction. The framework suggests a novel way of network classification: first, capture the latent affiliations of actors by extracting disease prediction based on network connectivity, and next, apply extant data mining techniques to classification based on the extracted prediction. The superiority of this framework over other representative relational learning methods has been verified with dengue prediction dengue data.

A. SVM

The two well-performing feature selection algorithms on the dataset are briefly outlined. Feature reduction applies a mapping of the multidimensional space into a space of lower dimensions. SVM is a linear transformation with linear orthonormal basis vectors; it can be expressed by a translation and rotation. Thereby improving the overall performance of classifier based intrusion detection systems. SVM is a linear transformation with linear ortho normal basis vectors; it can be expressed by a translation and rotation.

B. SMO

Classification is the type of Data mining, which deals with the problematic things by recognizing and detecting features of infection, among patients and forecast that which technique shows top performance, on the base of WEKA's outcome. Five techniques have been used in this paper. These techniques use Explorer interface and it depends on dissimilar techniques NB, REP Tree, RT, J48, and SMO.

- Better Accuracy rate.
- High Relevance Measure.
- Good classification accuracy.
- Online test data grouping process. High Similarity function provides types of dengue data classification.

4. Modules

- Data visualization and pre-processing.
- SMO feature selection algorithms.
- Feature reduction by SMO.

A. Module description

1) Data visualization and pre-processing

The Dengue Disease dataset is downloaded from the UCI (Unique Client Identifier) Machine Learning Repository website and saved as a document. This file is then imported into an Excel spreadsheet and therefore the values are saved with the corresponding attributes as column headers. The missing values are replaced with appropriate values. The ID of the patient cases doesn't contribute to the classifier performance. The algorithmic techniques applied for feature relevance analysis and classification are elaborately presented within the following sections.

2) SMO feature selection algorithms

The generic problem of supervised feature selection can be outlined as follows. Given a data set $\{(x_i, y_i)\}_{i=1}^n$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{1, 2, \dots, c\}$, we aim to find a feature subset of size m which contains the most informative features. The two well-performing feature selection algorithms on the WPDC dataset are briefly outlined below.

Mean and STD Score Filtering:

It is termed Univariate Mean and STD Score's ANOVA ranking. It is a supervised feature selection algorithm that processes the selection independently from the learning algorithm. It follows a filtering approach that ranks the input attributes according to their relevance. A cutting rule enables the selection of a subset of these attributes. It is required to define the target attribute which in this domain of research applies to the nature of the Dengue Disease (recurrent/non-recurrent) and the predictor attributes. After computing the Mean and STD Score for each feature, it selects the top- m ranked features with large scores.

3) Feature reduction by SMO

Feature reduction applies a mapping of the multidimensional space into a space of lower dimensions. Feature extraction includes features construction, space dimensionality reduction, sparse representations, and feature selection all these techniques are commonly used as pre-processing to machine learning and statistics tasks of prediction, including pattern recognition. Although such problems have been tackled by researchers for many years, there has been recently a renewed interest in feature extraction. The feature space having reduced features truly contributes to classification that cuts pre-processing costs and minimizes the effects of the 'peaking phenomenon' in classification. Thereby improving the overall performance of classifier based intrusion detection systems. The commonly used dimensionality reduction methods include supervised approaches such as Linear Discriminant Analysis (LDA), unsupervised ones such as SMO, and additional spectral and manifold learning methods. It converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. Consider the two dimensional cases then the basic principle of this transformation.

5. Architecture diagram

In figure 1, the dengue dataset is trained and loaded. The missing data are analyzed and filled with data visualization and pre-processing. Using the MLP input feature selection method ranking feature of the dataset is loaded and tested. The result is displayed using feature reduction using MLP.

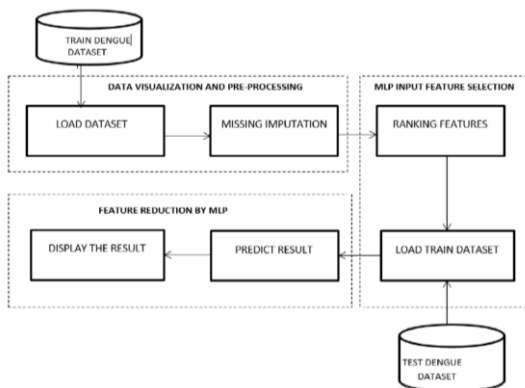


Fig. 1. Architecture of dengue prediction using MLP

6. Algorithm

The following steps illustrate SMO as follows, Consider the following,

Input: G = set of antigens to be recognized, n the amount of worst elements to pick for removal,

1. Create an initially random set of attributes, A for all antigens in G do.
2. Determine the expected with each attribute in an exceedingly.
3. Generate clones of a subset of the attribute in an exceedingly with the very best predicted.
4. The amount of clones for an attribute is proportional to its predicted.
5. Mutate attributes of those clones to the set A , and place a replica of the very best predicted.
6. Attribute in an exceedingly into the memory set, M .
7. Replace the n lowest predicted attributes in an exceedingly with new randomly generated attributes.
8. End

Output: M = set of memory attributes capable of classifying unseen patterns.

7. Result and discussion

The algorithms used here were applied to a dengue data set explained very well. So as to get better accuracy 10-fold cross validation was performed. For every classification we selected training and testing sample randomly from the bottom set to coach the model so test it so as to estimate the classification and accuracy measure for every classifier. The thrust classifications and accuracy utilized by us are:

Correctly Classified Accuracy:

It shows the accuracy percentage of test that's correctly classified.

Incorrectly Classified Accuracy:

It shows the accuracy percentage of test that's incorrectly classified.

Mean Absolute Error:

It shows the amount of errors to research algorithm classification accuracy.

Datasets:

Dataset could be a collection of knowledge or one statistical data where every attribute of knowledge represents variable and every instance has its own description. For prediction of dengue disease, we used dengue data set for prediction and classification of algorithms so as to match their accuracy using weka's three interfaces: Explorer, Experimenter and Knowledge Flow.

A description of dengue dataset:

The dataset utilized by us contains 18 attributes for dengue disease classification and accuracy. We've got applied different algorithms using WEKA data processing tool for our analysis purpose.

Table 1

Comparison of existing and the proposed system with their Accuracy, Sensitivity(SE), Specificity(SP) and Area Under Cover (AUC)

Method	Accuracy	Sensitivity (SE)	Specificity (SP)	Area Under Cover (AUC)
SMO Algorithm (Proposed System)	99.874	0.999	0.999	1
J48 Algorithm (Existing System)	96.201	0.958	0.966	0.993

Table 2, describes the attributes of knowledge set which are presented the file format of datasets used is Comma Separated Value CSV. Each attribute shows the current absent of dengue symptoms, number of days, date, number of WBC, number of platelets, pain and taste among patients in several cities and the way many days they suffers.

Table 2

Description of datasets attributes

Attributes	Description
P.I.D	Patient ID
Date of fever	Month
Residence	City
Days	No. of days
Current Temperature	Fever
WBC	No. of WBC
Severe Headache	Yes or No
Pain	Behind Eyes
Joint / Muscle pain	Yes or No
Metallic Taste	Yes or No
Appetite	Yes or No
Abdominal pain	Yes or No
Nausea/Vomiting	Yes or No
Diarrhoea	Yes or No
Haemoglobin	Haemoglobin Range
Haematocrit	Haematocrit Range
Platelets	No. of Platelets
Dengue	Yes or No

Figure 2, indicates the comparison made with the existing system which follows the decision tree, with the proposed work which is implemented successfully using Multi-Layer Perceptron (MLP) algorithm. The Specification, Serialization and Area Under Curve (AUC) of the algorithms are analysed and represented in the form of graph. This graph is drawn with the values taken in the table 1.

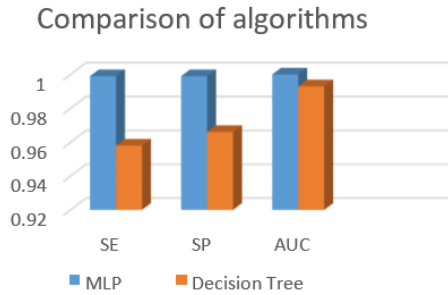


Fig. 2. Comparison of MLP (proposed algorithm) and decision tree (existing system) algorithms

8. Output

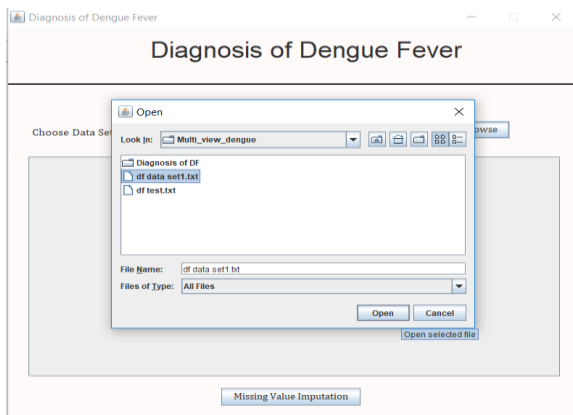
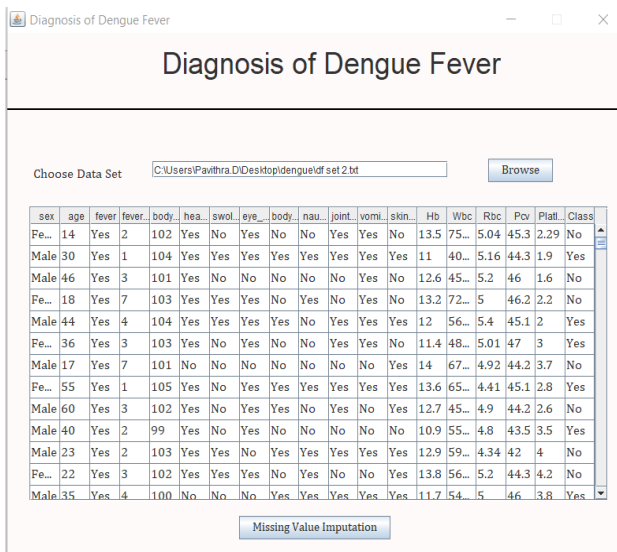
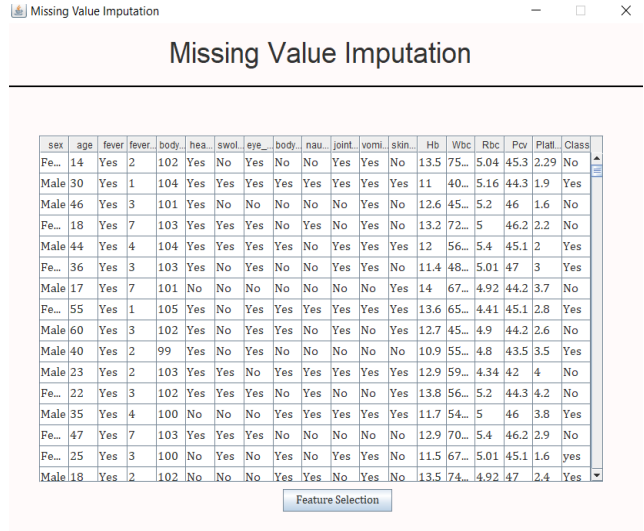


Fig. 3. Selection of the dengue dataset



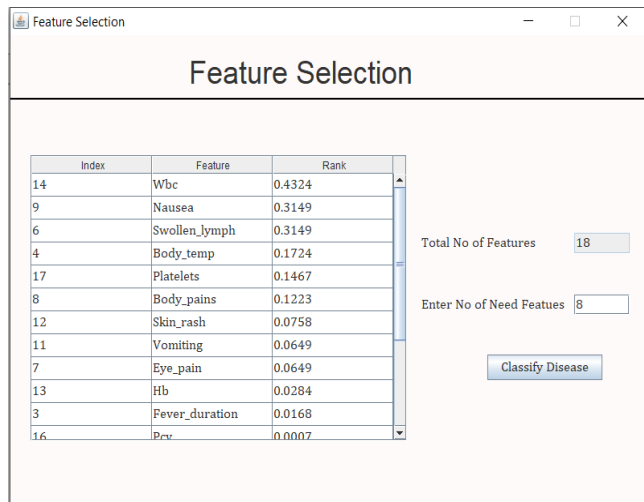
sex	age	fever	fever...	body...	hea...	swol...	eye...	body...	nau...	joint...	vomi...	skin...	Hb	Wbc	Rbc	Pcv	Plat...	Class
Fe...	14	Yes	2	102	Yes	No	Yes	No	Yes	Yes	No	13.5	75...	5.04	45.3	2.29	No	
Male	30	Yes	1	104	Yes	Yes	Yes	Yes	Yes	Yes	Yes	11	40...	5.16	44.3	1.9	Yes	
Male	46	Yes	3	101	Yes	No	No	No	No	No	No	12.6	45...	5.2	46	1.6	No	
Fe...	18	Yes	7	103	Yes	Yes	Yes	No	Yes	No	Yes	13.2	72...	5	46.2	2.2	No	
Male	44	Yes	4	104	Yes	Yes	Yes	Yes	No	Yes	Yes	12	56...	5.4	45.1	2	Yes	
Fe...	36	Yes	3	103	Yes	No	Yes	No	No	No	No	11.4	48...	5.01	47	3	Yes	
Male	17	Yes	7	101	No	No	No	No	No	No	No	14	67...	4.92	44.2	3.7	No	
Fe...	55	Yes	1	105	Yes	No	Yes	Yes	Yes	Yes	Yes	13.6	65...	4.41	45.1	2.8	Yes	
Male	60	Yes	3	102	Yes	No	Yes	No	Yes	No	Yes	12.7	45...	4.9	44.2	2.6	No	
Male	40	Yes	2	99	Yes	No	Yes	No	No	No	No	10.9	55...	4.8	43.5	3.5	Yes	
Male	23	Yes	2	103	Yes	Yes	No	Yes	Yes	Yes	Yes	12.9	59...	4.34	42	4	No	
Fe...	22	Yes	3	102	Yes	Yes	Yes	No	Yes	No	No	13.8	56...	5.2	44.3	4.2	No	
Male	35	Yes	4	100	No	No	No	Yes	Yes	Yes	Yes	11.7	54...	5	46	3.8	Yes	
Fe...	47	Yes	7	103	Yes	Yes	Yes	No	No	No	No	12.9	70...	5.4	46.2	2.9	No	
Fe...	25	Yes	3	100	No	Yes	No	Yes	No	Yes	No	11.5	67...	5.01	45.1	1.6	Yes	
Male	18	Yes	2	102	No	No	No	Yes	Yes	No	Yes	13.5	74...	4.92	47	2.4	Yes	

Fig. 4. Raw dengue dataset



sex	age	fever	fever...	body...	hea...	swol...	eye...	body...	nau...	joint...	vomi...	skin...	Hb	Wbc	Rbc	Pcv	Plat...	Class
Fe...	14	Yes	2	102	Yes	No	Yes	No	No	Yes	No	13.5	75...	5.04	45.3	2.29	No	
Male	30	Yes	1	104	Yes	Yes	Yes	Yes	Yes	Yes	Yes	11	40...	5.16	44.3	1.9	Yes	
Male	46	Yes	3	101	Yes	No	No	No	No	No	Yes	12.6	45...	5.2	46	1.6	No	
Fe...	18	Yes	7	103	Yes	Yes	Yes	No	Yes	No	Yes	13.2	72...	5	46.2	2.2	No	
Male	44	Yes	4	104	Yes	Yes	Yes	No	Yes	Yes	Yes	12	56...	5.4	45.1	2	Yes	
Fe...	36	Yes	3	103	Yes	No	Yes	No	No	No	No	11.4	48...	5.01	47	3	Yes	
Male	17	Yes	7	101	No	No	No	No	No	No	No	14	67...	4.92	44.2	3.7	No	
Fe...	55	Yes	1	105	Yes	No	Yes	Yes	Yes	Yes	Yes	13.6	65...	4.41	45.1	2.8	Yes	
Male	60	Yes	3	102	Yes	No	Yes	No	Yes	No	Yes	12.7	45...	4.9	44.2	2.6	No	
Male	40	Yes	2	99	Yes	No	Yes	No	No	No	No	10.9	55...	4.8	43.5	3.5	Yes	
Male	23	Yes	2	103	Yes	Yes	No	Yes	Yes	Yes	Yes	12.9	59...	4.34	42	4	No	
Fe...	22	Yes	3	102	Yes	Yes	Yes	No	Yes	No	No	13.8	56...	5.2	44.3	4.2	No	
Male	35	Yes	4	100	No	No	No	Yes	Yes	Yes	Yes	11.7	54...	5	46	3.8	Yes	
Fe...	47	Yes	7	103	Yes	Yes	Yes	No	No	No	No	12.9	70...	5.4	46.2	2.9	No	
Fe...	25	Yes	3	100	No	Yes	No	Yes	No	Yes	No	11.5	67...	5.01	45.1	1.6	Yes	
Male	18	Yes	2	102	No	No	No	Yes	Yes	No	Yes	13.5	74...	4.92	47	2.4	Yes	

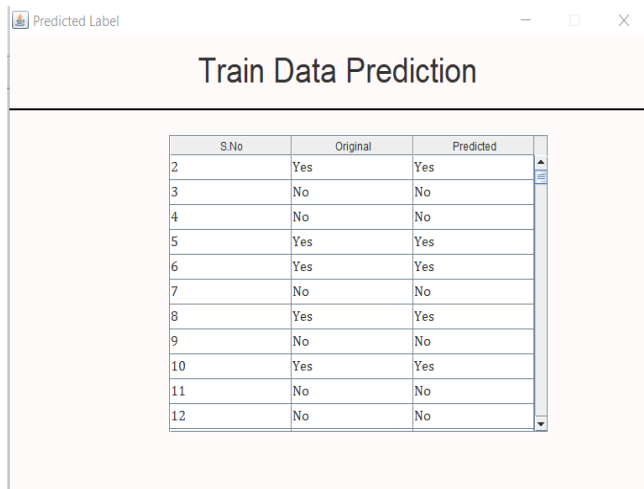
Fig. 5. Finding the missing value using K - means algorithm



Index	Feature	Rank
14	Wbc	0.4324
9	Nausea	0.3149
6	Swollen_lymph	0.3149
4	Body_temp	0.1724
17	Platelets	0.1467
8	Body_pains	0.1223
12	Skin_rash	0.0758
11	Vomiting	0.0649
7	Eye_pain	0.0649
13	Hb	0.0284
3	Fever_duration	0.0168
16	Pcv	0.0007

Total No of Features: 18
 Enter No of Need Features: 8
 Classify Disease

Fig. 6. Feature selection



S.No	Original	Predicted
2	Yes	Yes
3	No	No
4	No	No
5	Yes	Yes
6	Yes	Yes
7	No	No
8	Yes	Yes
9	No	No
10	Yes	Yes
11	No	No
12	No	No

Fig. 7. Train data prediction

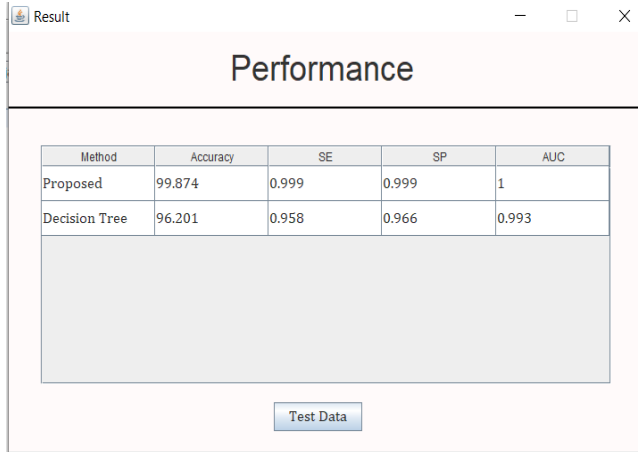


Fig. 8. Performance comparison

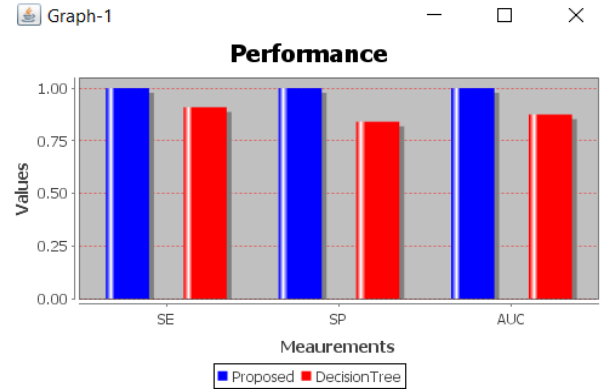


Fig. 11. Comparison of existing and proposed system

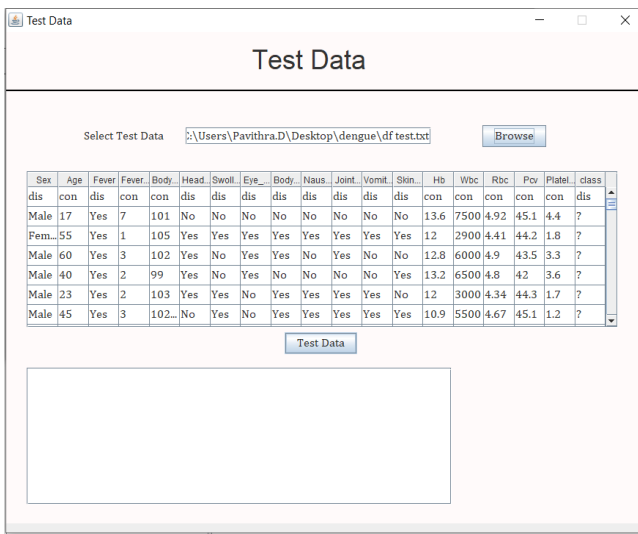


Fig. 9. Test data of dengue

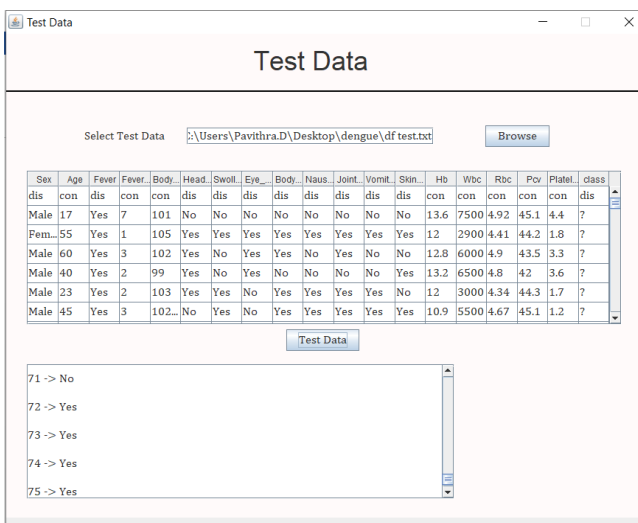


Fig. 10. Dengue test data prediction

9. Conclusion

The research test different algorithms. The result of the research focused on the correctness of the algorithms in the training. It depended on the WDBC data set. The test result shows that SMO is the best algorithm. The best way was when the research removed the sample for the missing value in training for SMO. However, the Random Tree result kept better correctness when keeps the sample for the missing value. The research undertook an experiment on the application of various data mining algorithms to predict the dengue and to compare the best method of prediction.

The research results do not present dramatic differences in the prediction when using different classification algorithms in data mining. The experiment can serve as an important tool for physicians to predict risky cases in the practice and advise accordingly. The model from the classification will be able to answer more complex queries in the prediction of dengue diseases. The predictive accuracy determined by the SMO algorithm suggests that the parameters used are reliable indicators to predict the presence of dengue diseases.

Acknowledgement

We would like to express our sincere gratitude to several individuals and organizations for supporting me throughout my Graduate study. First, we wish to express our sincere gratitude to my supervisor, Associate Professor Dr. V. Janani, for her enthusiasm, patience, insightful comments, helpful information, practical advice and unceasing ideas that have helped us tremendously at all times in our research and writing of this paper. Her immense knowledge, profound experience and professional expertise in Data Mining has enabled us to complete this research successfully. Without her support and guidance, this project would not have been possible. I also wish to express my sincere thanks to the Adhiyamaan College of Engineering for accepting our project into the graduate program.

References

- [1] Blackburn G L, Wang K. A. (2007) "Dietary fat reduction and Dengue Disease outcome: results from the Women's Intervention Nutrition Study

- (WINS)", International Journal of Computer Science and Engineering, vol. 32, pp. 512.
- [2] Boffetta P, Hashibe M (2006). "The burden of cancer attributable to alcohol drinking", Journal of Cancer Research and Treatment, vol. 90, pp. 119.
- [3] Boris Pasche (2010). "Cancer Genetics", (Cancer Treatment and Research). Berlin: Springer. pp. 19–20.
- [4] Collaborative Group on Hormonal Factors in Dengue Disease (2002), "Dengue Disease and breastfeeding", International Journal for Cancer Research and Treatment, vol. 15.
- [5] Delen D, Walker G, (2005), "Predicting Dengue Disease survivability: a comparison of three data mining methods", Artificial Intelligence in Medicine, vol. 34, pp. 113-127.
- [6] Ferro, Roberto (2012), "Pesticides and Dengue Disease", International Journal for Cancer Research and Treatment, vol.76.
- [7] Gage M, Wattendorf D (2012), "Translational advances regarding hereditary Dengue Disease syndromes", International Journal of Computer Science and Engineering, vol. 90.
- [8] Johnson KC, Miller AB, (2009). "Active smoking and secondhand smoke increase Dengue Disease risk: the report of the Canadian Expert Panel on Tobacco Smoke and Dengue Disease Risk (2009)," Journal of Computer Science and Engineering. Vol.69, pages 198-199.
- [9] Gupta S.; Kumar D., Sharma A, (2011), "Data Mining Classification Techniques Applied for Dengue Disease Diagnosis and Prognosis", Journal of Computer Science and Engineering, vol. 23, pp. 1191-1193.