iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** IX **Month of publication:** September 2017

DOI: http://doi.org/10.22214/ijraset.2017.9027

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue IX, September 2017- Available at www.ijraset.com

Predictive Analysis of Different Classification Techniques of I Virus - H4n2 Influenza Using Weka

Dr.S.Kavitha[1], Dr.M.Hanumanthappa[2]
[1]*Research Scholar, Rayalaseema University, Kurnool*
[2]*Professor, Bangalore University,Bangalore*

Abstract: ***Data mining is a powerful tool to be used in real time applications as its artificial intelligence nature. Data mining is highly used in medical domain as it helps in making better predictions and supports in decision making. It also supports physicians in developing better diagnostic treatments.'I' Virus is the Influenza virus of type A. The influenza virus consists of eight RNA strand. The h and N full form is hemagglutinin (H) and neuraminidase (N) respectively. There are 18 different H antigens (H1 - H18) and N antigens have 11 different N (N1 to N11). Variations in the chemical character will give different influenza virus. Symptoms of influenza are respiratory problems and muscular aches. In this work various symptoms for I virus causes are discussed. The dataset arff file with seven attributes for sample test cases are used. Weka has a large collection of learning algorithms, most of which are batch-based and operate on data held in main memory. 64 bit VMs extend the amount of data that these methods can operate, but available RAM still limits the amount of data that can be processed. In this work data mining technique used for classification of diseases such as dengue, diabetes and cancer in bioinformatics research. In the proposed approach we have used WEKA with 10 cross validation to evaluate data and compare results. In this paper we have firstly classified the I-virus data set and then compared the different data mining techniques in weka through Explorer, knowledge flow and Experimenter***
interfaces. Furthermore in order to validate our approach we have used I-Virus dataset with 75 instances and 12 attributes to determine the prediction of disease and their accuracy using classifications of different algorithms to find out the best performance. The main objective of this paper is to classify data and assist the users in extracting useful information from data and easily identify a suitable algorithm for accurate predictive model from it. Using I-virus email dataset Preprocessing and classification is workout in Weka Explorer. Multiple classification Techniques can be examined to get better result. By analyzing all these techniques, the Random Tree with Partition Membership Filter gives better Performance than others. Here Weka tool is used as a software tool to analyze result.
Key words: Data Mining, h4N2 influenza, I-Virus,Weka,Classifier,Clustering

I. INTRODUCTION

The I virus (H4N2 influenza virus) basic characteristic is changing physiology. The influenza virus will not transfer to humans directly from the environment. Rarely, it is observed in human being. H4N2 influenza virus develops through 3 stages. In the first stage, anterior pituitary glands grow very fast and the epithelial cells forming mature glands(8). In the second stage epithelial cells will become flatter and lumen enlarges abnormally. Influence of adreno cortical hormones, full secretory activity. The third stage or final stage of H4N2 influenza virus is the development stage where resting glands in the formed ducts. It will be like a Mammary Tumour Virus (MTV).

Figure 1. I virus (H4N2 influenza virus)

Influenza virus is classified into 3 types: Influenza virus A,Influenza virus B,Influenza virus C Influenza A virus mainly affects birds and rarely in mammels. Viruses are transmitted from wild aquatic birds to domestic poultry, and then to human influenza pandemics. Rarely, it leads to death.The fundamental characteristics of the I virus are changing physiology. Usually I virus will not transfer to humans through infection or from the environment. A comprehensive study of virus-induced cell alterations must therefore take into account the intimate biological modifications which normally arise in the individual cell as well as in its immediate environment.

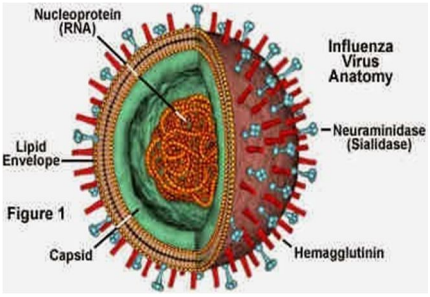


Figure 2. Structure of I virus (H4N2 influenza virus)

The I virus technically called as Hunchback (Kyphosis) Adreno Cortical, h4N2 influenza(8). If virus is injected to human and soon victim starts to see the symptoms like hair fall, weaking of muscles, loss of weight, spinal problems. When patient enters into the final stage, he gets hunch back. At a later stage, he gets treatment by a doctor and slowly he will be cured by medicine and exercise.

A. I Virus (H4n2) Dataset

A type of Influenza virus h4N2 is called as I virus, a novel virus invented or man made virus. This type influenza virus is rarely seen infect human being. It is mostly observed to affect the birds and animals. It is labelled as 2 strands of hemagglutinin (H) and 4 strands of neuraminidase (N). H antigens vary from H1 to H18 and N antigens varies from N1 to N11, totally these 18 H and 11 N various combinations give rise to different influenza virus based on chemical charteristics. The I virus only effect physically but not mentally. This virus will damage all the physical body cell including tissues the symptoms of this virus are bend in spinal cord, Hair Fall, Blisters through out the skin, weight loss, Getting older faster than the normal person.

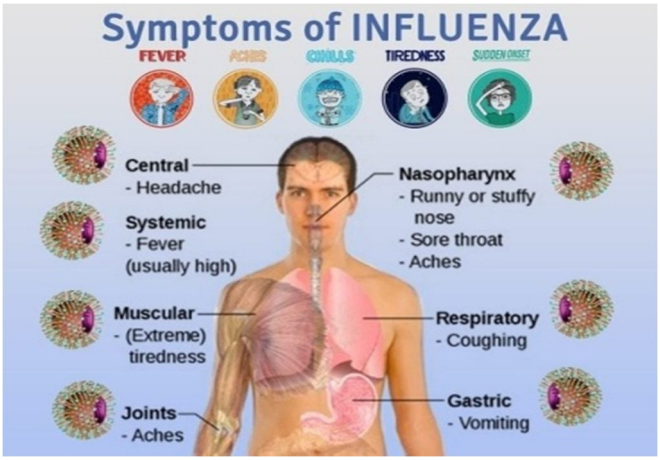


Figure 3. Symptoms of Influenza

II. METHODOLOGY

Weka has a large collection of learning algorithms, most of which are batch-based and operate on data held in main memory. 64 bit VMs extend the amount of data that these methods can operate on (within the limits of computational complexity), but available RAM still limits the amount of data that can be processed(7). There are a number of approaches to consider when data is too large to fit into main memory.

A. Characteristics of Data

The first thing to consider is the nature of the data itself. If there are columns that can be safely eliminated apriori (i.e. identifiers, dates, constant fields etc.) then this should be done in a preprocessing step. If the data contains many zeros then Weka has a sparse

data format that can save a lot of memory(6). All algorithms in Weka can take advantage of the memory savings afforded by the sparse format; some algorithms can take advantage of the sparsity to speed up computation.

B. Streaming incremenatal learners

Quite a few learning algorithms in Weka that can be trained incrementally, one data row at a time. These methods generally have runtime that is linear in the number of rows and fields in the data and only require the current data row to be present in main memory. Because of this, they can be used to process a (potentially) infinite number of rows. Furthermore, due to their incremental nature, they are "anytime" algorithms that can be used for prediction at any stage. Their current state can be saved and training resumed at a later time. Algorithms included in Weka that fall into this category are:

1) naive Bayes
2) naive Bayes multinomial (naive Bayes for text categorization)
3) DMNBtext (discriminitive multinomial naive bayes for text categoriziation)
4) AODE and AODEsr (averaged one dependence estimators)
5) SPegasos (the Pegasos stochasitc gradient descent algorithm for learning linear support vector machines and logistic regression for binary class problems)
6) SGD (stochastic gradient descent for linear regression, binary class logistic regression and linear support vector machines)
7) IB1, IBk and KStar (nearest neighbor learners for classification and regression using a sliding window on the data)
8) locally weighted learning (locally weighted models using a sliding window on the data)
9) RacedIncrementalLogitBoost (ensembles of boosted base learners applied to data "chunks")
10) Cobweb (incremental clustering)

Furthermore, Weka >=3.7.2 includes a package that makes the MOA (Massive Online Analysis) data stream learners available in Weka. This adds Hoeffding incremental decision trees, Hoeffding trees with naive Bayes models at the leaves, Hoeffding option trees(5), bagging, boosting and adaptive variants thereof. MOA also includes routines for automatic memory management that can prevent a learned model from exceeding a user-specified maximum memory constraint.

Weka's "Explorer" environment is batch-based (regardless of whether the algorithm used is incremental or not) and loads the data into main memory. To take advantage of incremental learning either the command line interface or the graphical Knowledge Flow interface (using "instance" connections) can be used.

Although not actually a streaming algorithm, the FPGrowth association rule learner can (as of Weka 3.7.2) operate off of data stored on disk rather than loaded into main memory. It requires only two passes over the data in order to build its extended prefix tree structure. This prefix tree is held in main memory however, so despite the compression it usually achieves(4) (which is partly dependent on data characteristics such as sparseness), there will be a limit to the number of rows/transactions it can handle.

C. Sampling

Quite often a carefully selected subsample of the data can be used to train a highly accurate model that usually has performance close to what would be achieved if the full data set could be processed. How is this possible? When plotting the accuracy of model as a function of the size of the data used to train it, the curve for a given scheme will flatten out after a certain amount of data has been processed. Learning schemes with high bias, such as naive Bayes and linear models will achieve most of their accuracy with a very small amount of data. Processing more data is usually not worth the effort (diminishing returns). So, selecting a suitably sized subsample of the data is a viable approach in these cases. Weka includes a pre-processing filter that can perform reservoir sampling (uniformly random sampling from streamed data when the number of rows is not known in advance).

D. Handling the classifiers

One simple approach to handling large data sets is to create an ensemble of classifiers by splitting the data into a number of chunks, where each chunk is capable of being loaded and processed in main memory. A separate classifier can be learned and then saved for each data chunk. This process can naturally be distributed across multiple machines (if available). The resulting ensemble of classifiers can be used for prediction by combining the predictions from the individual classifiers. Simple voting (classification) or averaging (probabilities for classification or predicted target values for regression) works well.

Below is the pedagogy of stepwise learning which will help you to understand the concepts in a better & concrete manner:

1) *Step 1 : What is Weka and Why to use it?:* Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. You might want to have a look at this video from Brandon Weinberg. This video will give you considerable insight to this amazing tool. You might not understand everything through this video but will certainly get a hang of things.
2) *Step 2 : Setting up Machine:* Now, that we are acquainted with Weka, we can proceed to the next stage. To know more about the tool and the people behind its success, you can have a look at this site on Project Weka. Moreover, you can also download the software and get the latest version for your system from this link.
3) *Step 3 : Learning the Basics of Weka:* The best way of getting started with Weka is using MOOC offered by University of Waikato. Data Mining with Weka is a well reputed course, but it isn't available around the year. Yet, not to worry, in such cases one can access the course videos from this Youtube Channel. The official link of this course can be viewed here. The Data Sets which will be discussed(3) in here can be downloaded from this link. The page has further links to data sets. Weka uses data in ARFF format. In case data is not in ARFF format, you can convert it from CSV to ARFF format by taking help from this video.
4) *Step 4 : Data Sets:* Having tried our hands at Data sets provided by the course coordinators, we will try our hands on a fresh data set from Kaggle. Since the format would be of .csv, convert it to ARFF format, so that we can read it onto the Weka interface. After having done these courses, once has attained enough skills to start working and analyzing data sets using Weka GUI(1) . Those who visited the MOOC link would have the seen the course 'More Data mining with Weka'. How to Run Your First Classifier in Weka Weka makes learning applied machine learning easy, efficient, and fun. It is a GUI tool that allows you to load datasets, run algorithms and design and run experiments with results statistically robust enough to publish.

III. RESULTS AND DISCUSSIONS

A. *Classifiers*

Classifier	ZeroR	Decision Table	JRIP	PART -M 2 -C 0.25 -Q 1
Correctly Classified Instances	54.386 %	73.6842 %	73.6842 %	61.4035 %
Incorrectly Classified Instances	45.614 %	26.3158 %	26.3158 %	38.5965 %
Kappa statistic	0	0.4968	0.4968	0.2335
Mean absolute error	0.3566	0.2539	0.2539	0.2797
Root mean squared error	0.42	0.3714	0.369	0.4549
Relative absolute error	100 %	79.8108 %	71.22 %	78.4384 %
Root relative squared error	100 %	88.4373 %	87.8458 %	108.3136 %

Table 1. Comparative study of varios classifiers for I-Virus dataset

Weighted Avg	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC ARea
ZeroR	0.544	0.544	.296	0.544	0.383	0.0	0.430	0.433
Decision Table	0.737	0.224	0.732	0.737	0.726	0.511	0.665	0.595
JRIP	0.737	0.224	0.732	0.737	0.726	0.511	0.659	0.577
PART -M 2 -C 0.25 -Q 1	0.614	0.387	0.591	0.614	0.601	0.232	0.641	0.580

Table 2. Comparative study of Weighted Average for various measures

B. *Study on various clusters with time taken to build and Incorrectly clustered instances*

1) EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100

Time taken to build model (full training data) : 0.7 seconds
Log likelihood: 1.70118

Incorrectly clustered instances : 26.0 44.0678 %

2) Canopy -N -1 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t2 -1.0 -t1 -1.25 -S 1

Time taken to build model (full training data) : 0 seconds
Incorrectly clustered instances : 48.0 81.3559 %

3) Cobweb -A 1.0 -C 0.0028209479177387815 -S 42

Time taken to build model (full training data) : 0.02 seconds
Incorrectly clustered instances : 55.0 93.2203 %

4) Farthest First -N 2 -S 1

Incorrectly clustered instances : 26.0 44.0678 %

5) kMeans

Time taken to build model (full training data) : 0 seconds
Incorrectly clustered instances : 24.0 40.678 %

6) HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"

Time taken to build model (full training data) : 0.02 seconds
Incorrectly clustered instances : 28.0 47.4576 %

7) MakeDensityBasedClusterer -M 1.0E-6 -W weka.clusterers.SimpleKMeans -- -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

kMeans
Time taken to build model (full training data) : 0.01 seconds
Incorrectly clustered instances : 24.0 40.678 %

8) MakeDensityBasedClusterer -M 1.0E-6 -W weka.clusterers.SimpleKMeans -- -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

239 240.9259 237.375
Time taken to build model (full training data) : 0 seconds
Incorrectly clustered instances : 24.0 40.678 %

9) SimpleKMeans -init 0 -max-candidates 100 -periodic-pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -num-slots 1 -S 10

Number of iterations: 5
Time taken to build model (full training data) : 0 seconds
Incorrectly clustered instances : 24.0 40.678 %

IV. VISUALIZATIONS

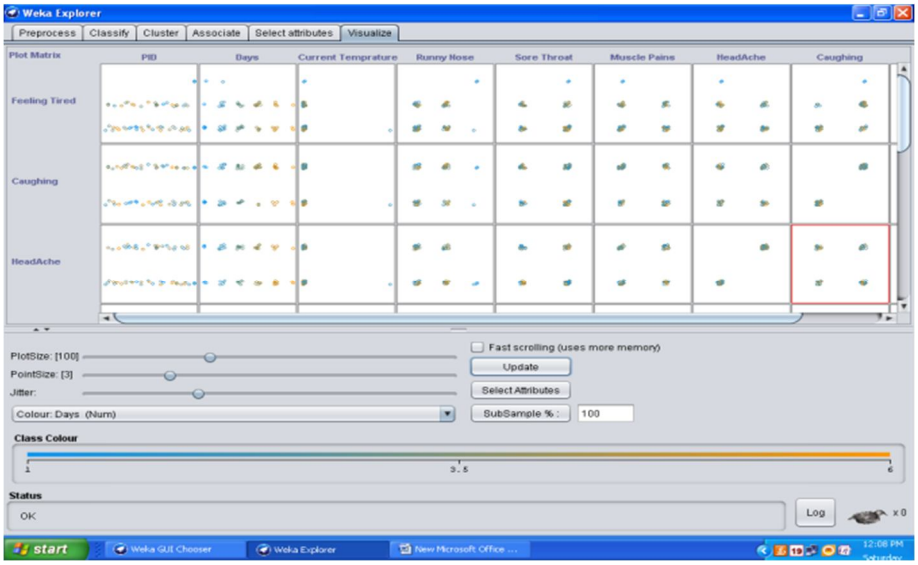


Figure 4. Screenshot view of the number of days Fever

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue IX, September 2017- Available at www.ijraset.com

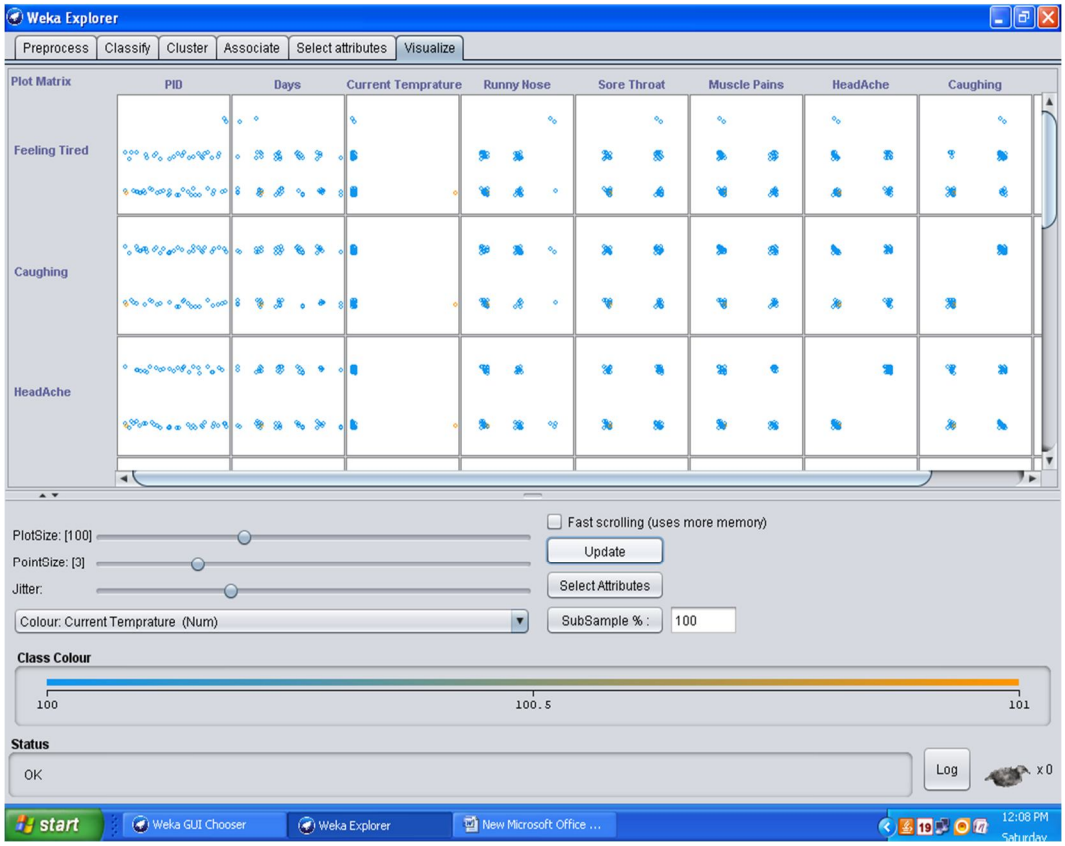


Figure 5. Screenshot view of current temparature

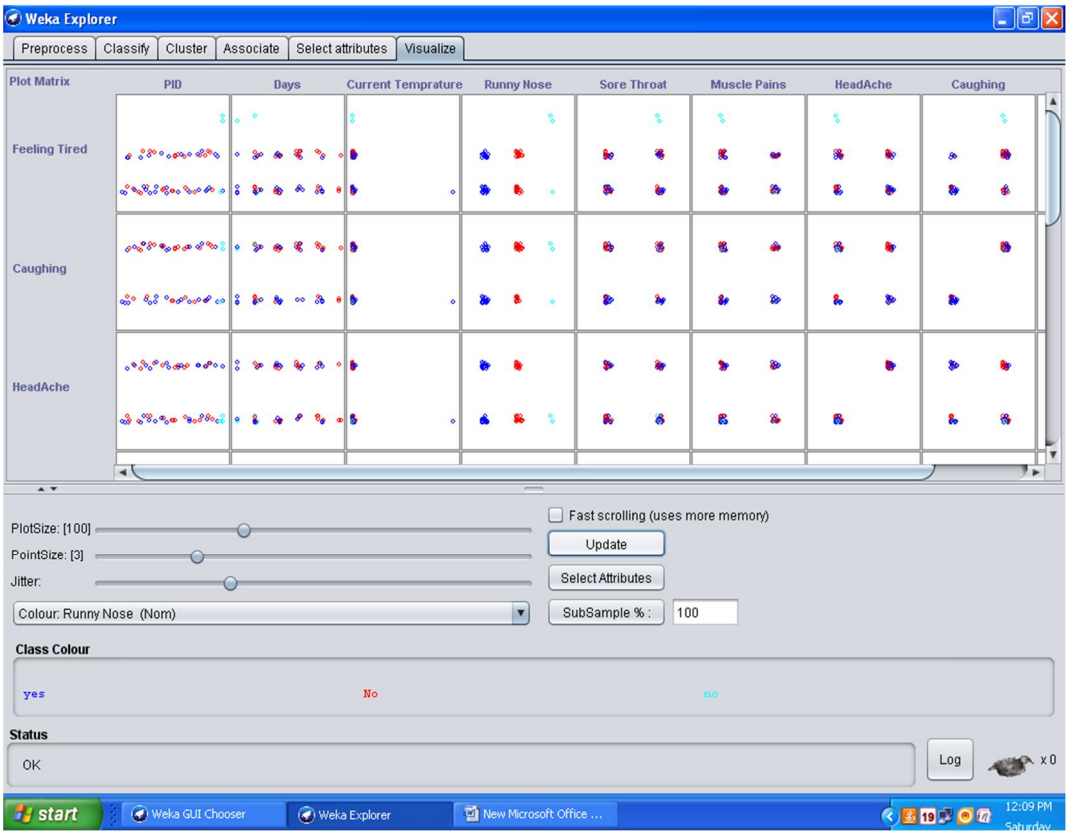


Figure 6. Screenshot view of Runny Nose

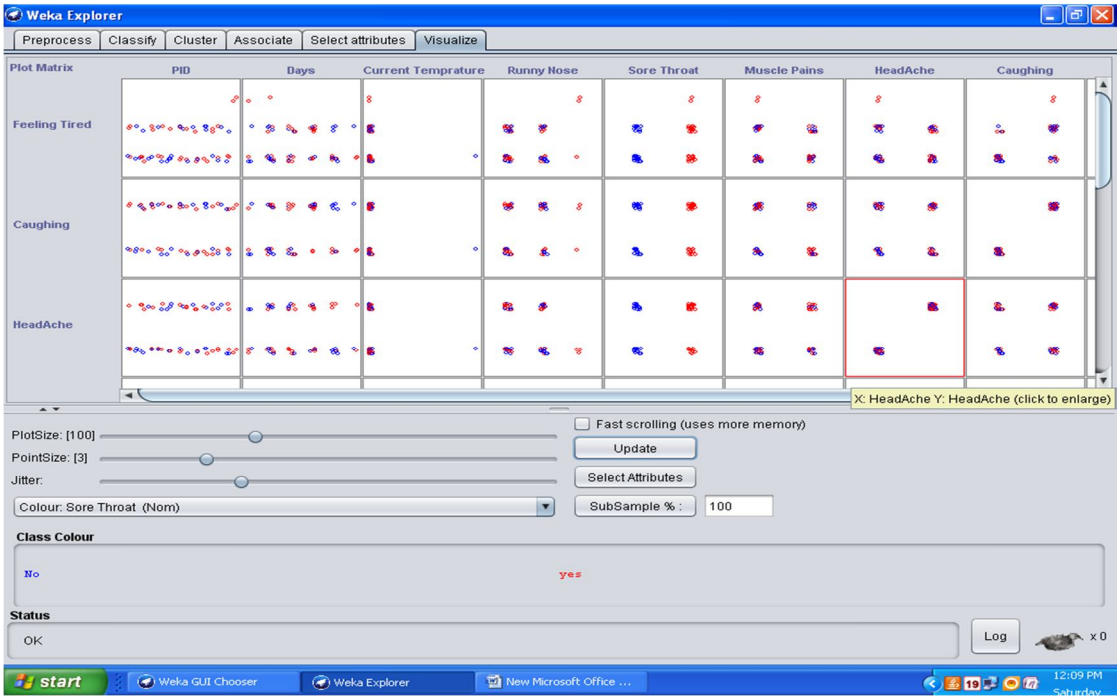


Figure 7. Screenshot view of Sore throat

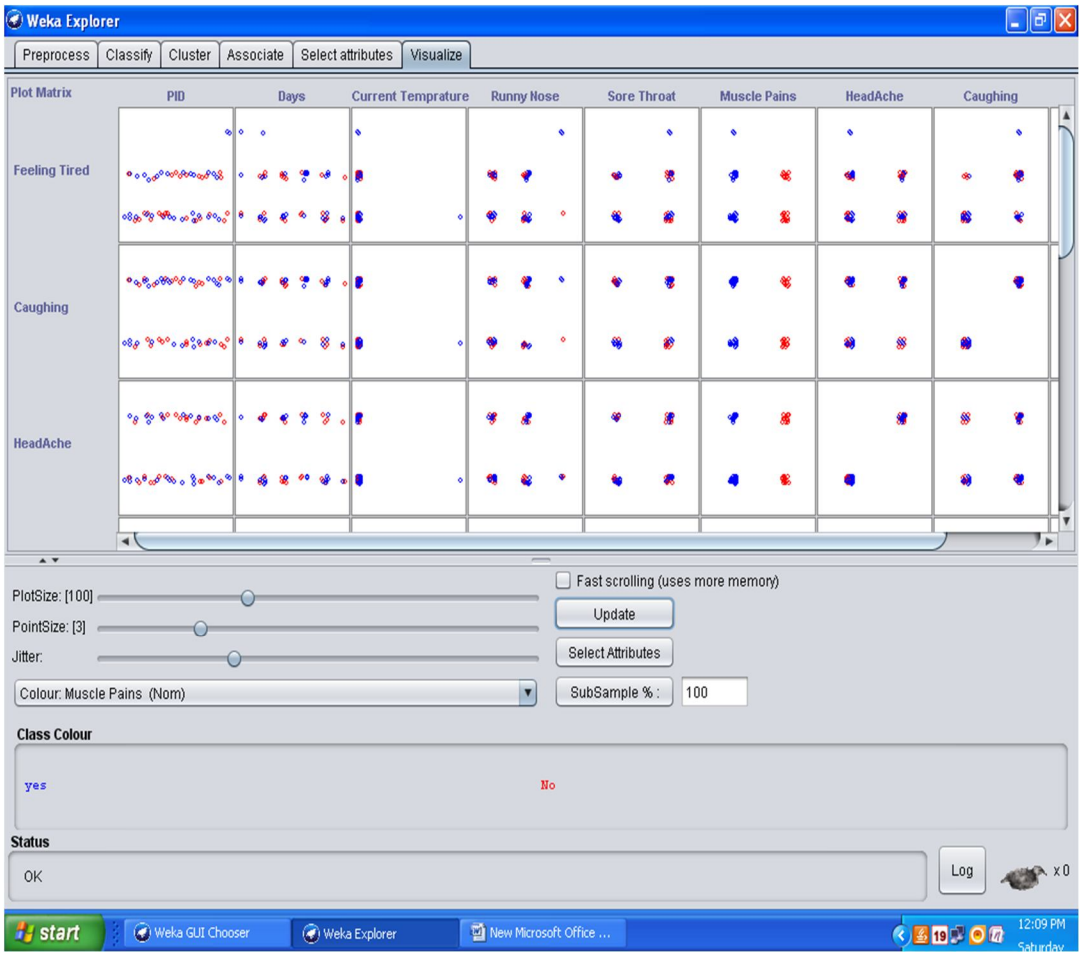


Figure 8. Screenshot view of Muscle pains

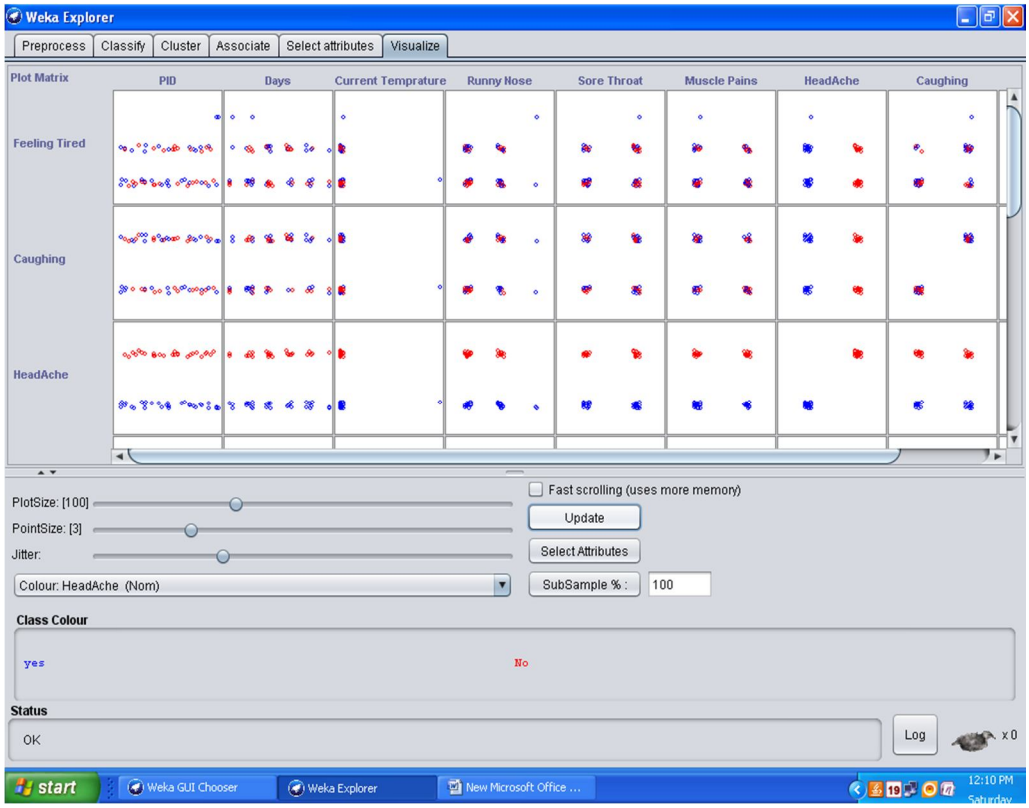


Figure 9. Screenshot view of Headache

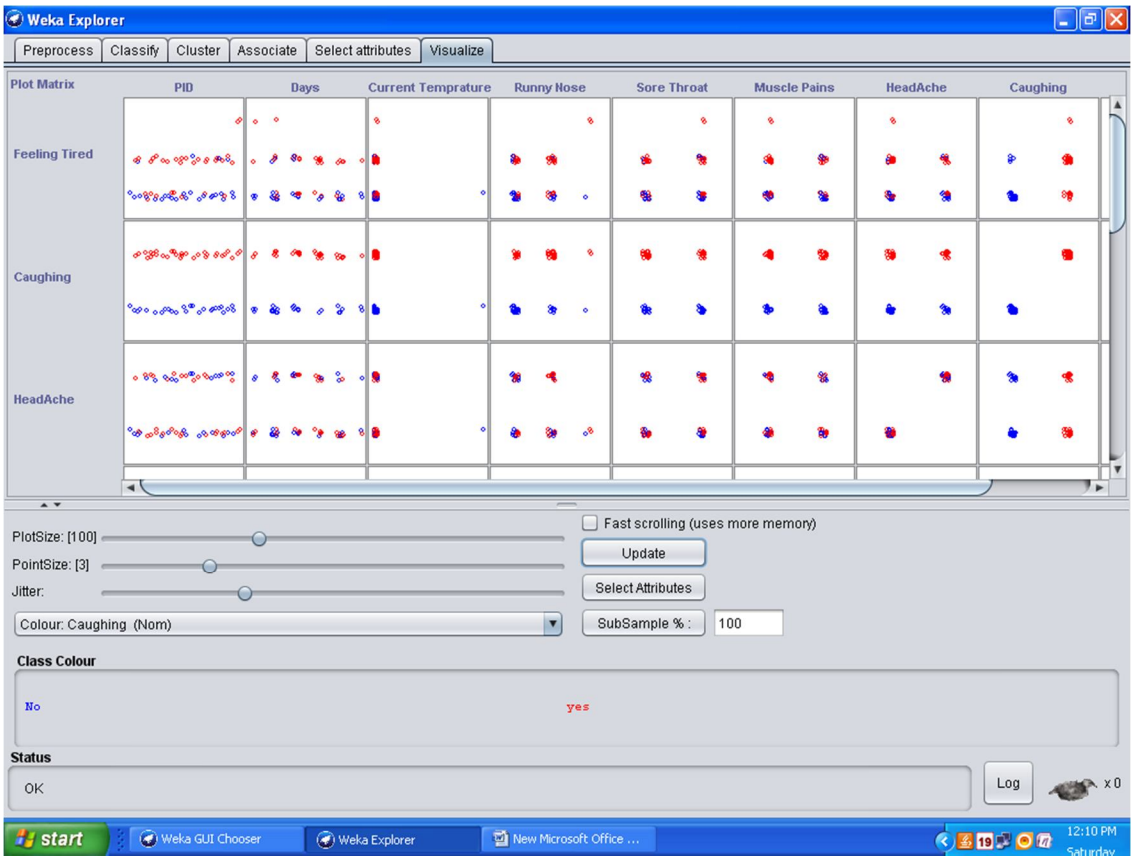


Figure 10. Screenshot view of Caughing

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor:6.887
Volume 5 Issue IX, September 2017- Available at www.ijraset.com

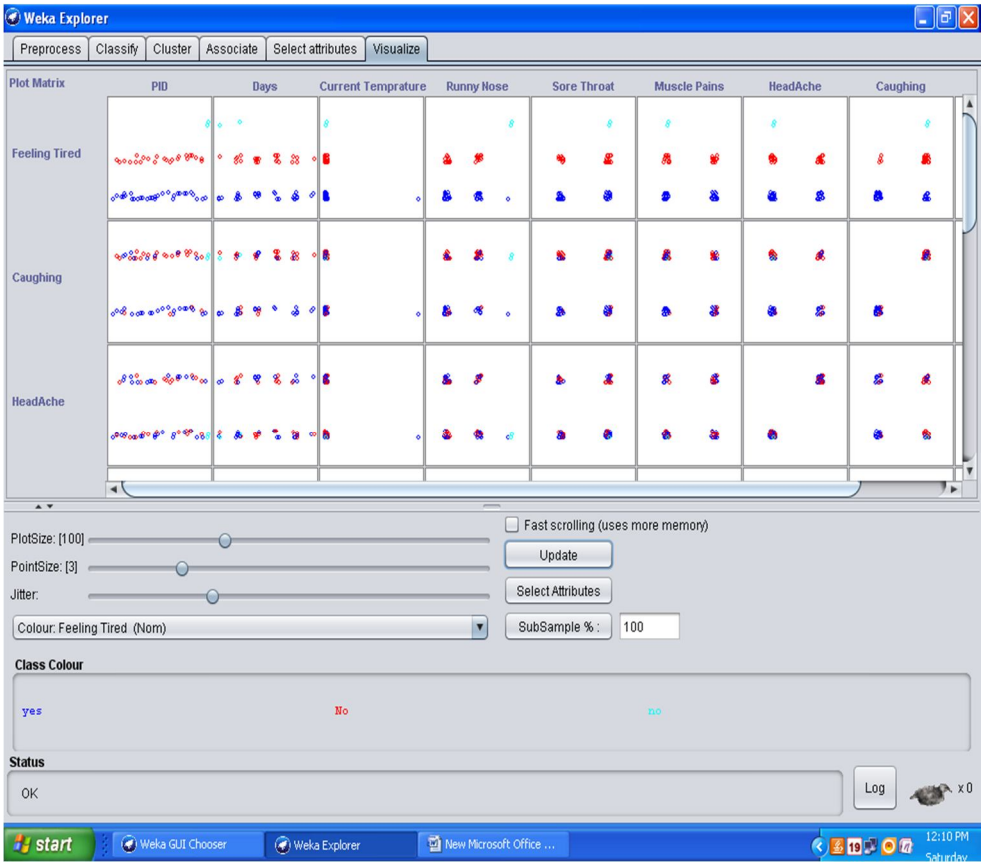


Figure 11. Screenshot view of feeling tired

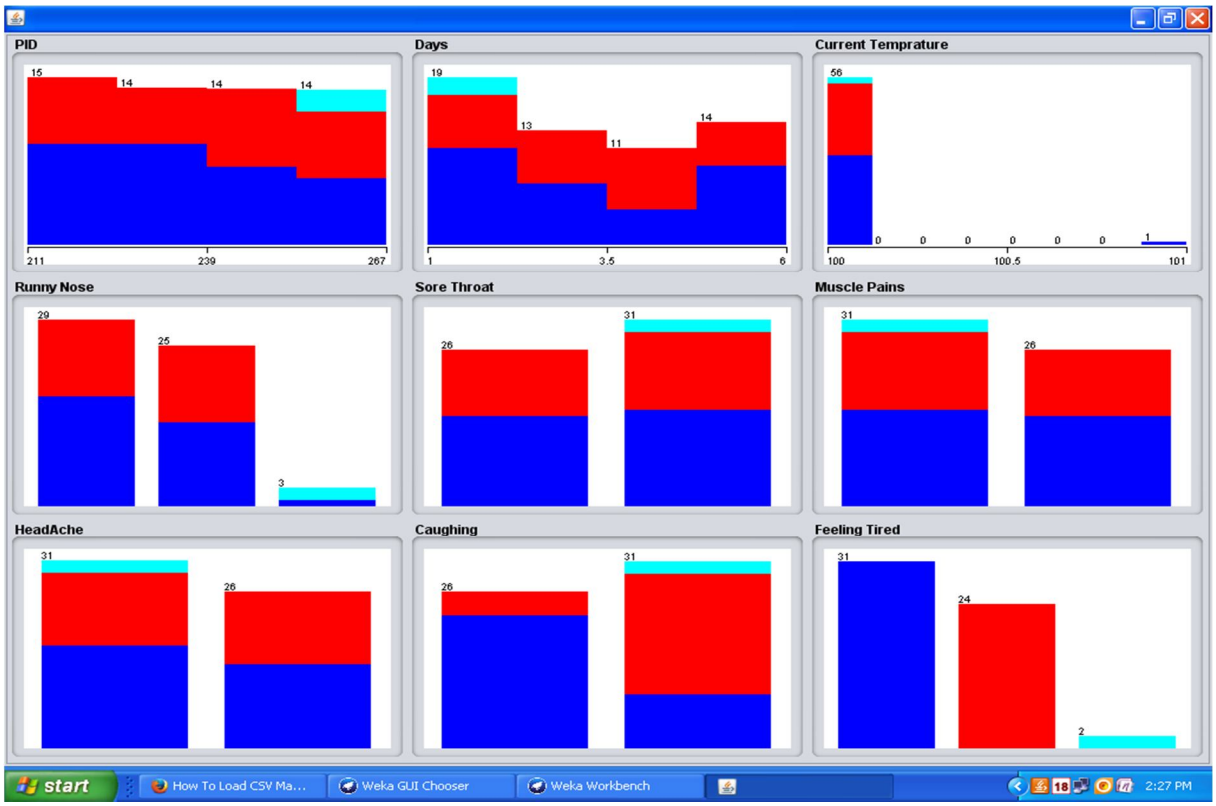


Figure 12. Screenshot view of attributes

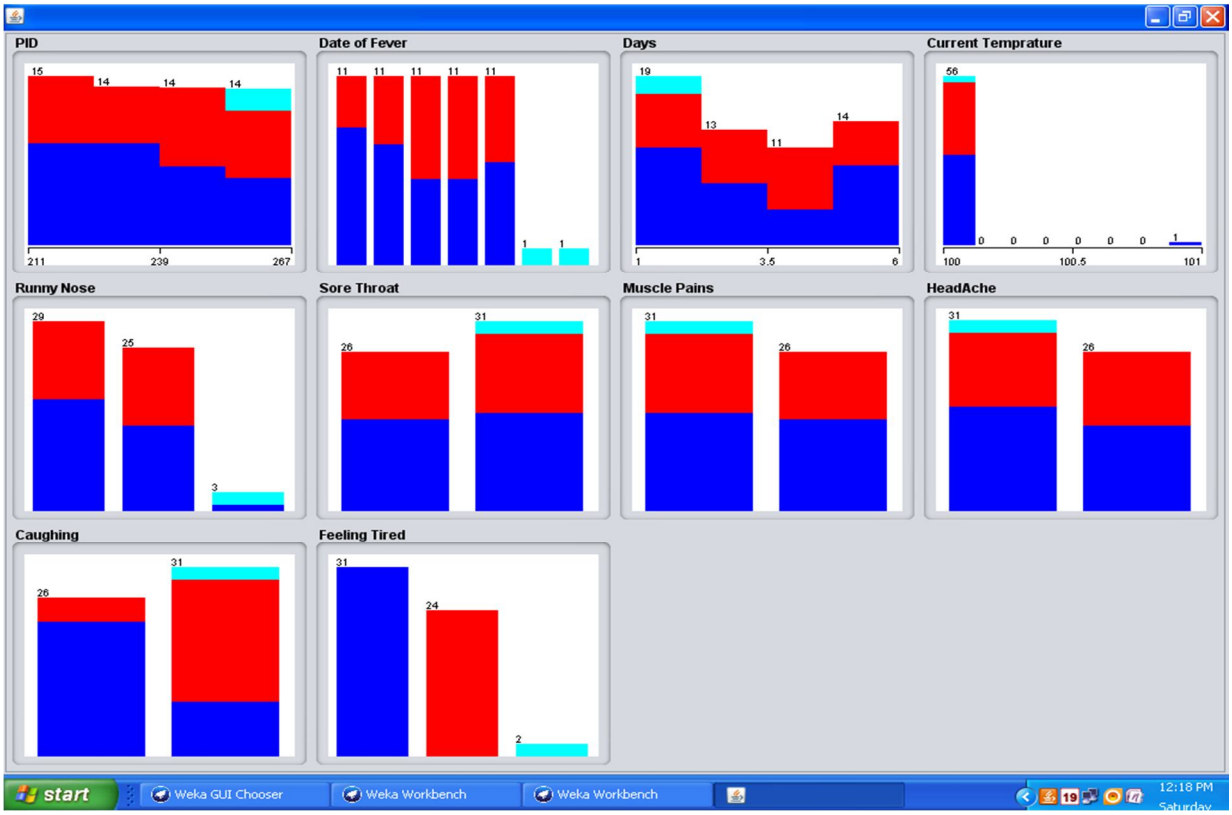


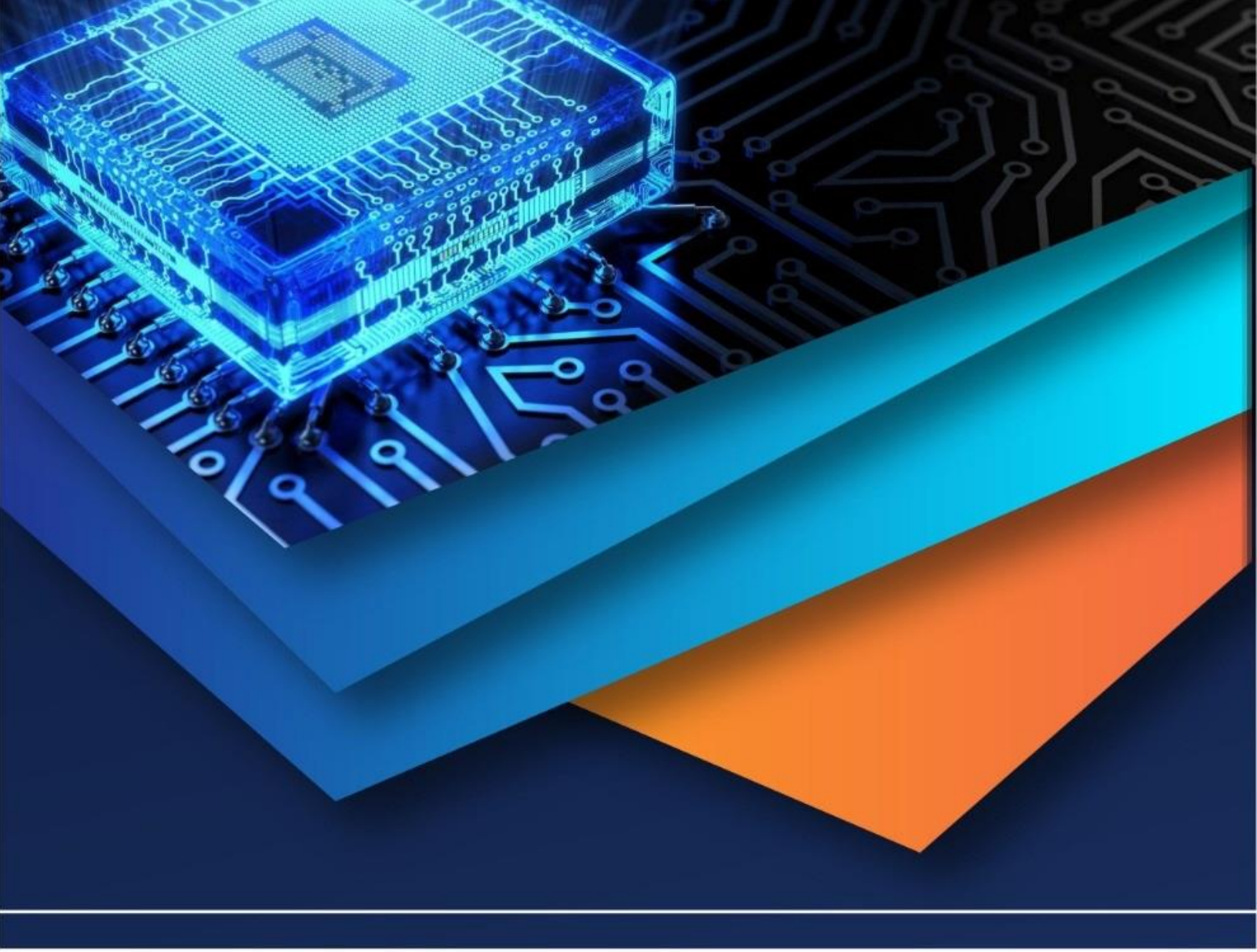
Figure 13. Screenshot view of I-virus attibutes with instances

V. CONCLUSION

The main aim of this paper is to analyse the study on I Virus - h4N2 influenza using WEKA data mining tool. In this paper we have used four classifiers ZeroR,Decision Table, JRIP, PART -M 2 -C 0.25 -Q 1.Then these classifiers were tested using WEKA tool to analyze classifier accuracy. After running these classifiers the outputs were compared on the basis of accuracy achieved. These classifiers accuracy to each other on the basis of correctly classified instances, Incorrectly classified instances, Kappa statistic, Mean Absolute error, root mean squared error, Relative absolute error, Root relative squared error. According to the experiment results of our study it was predicted that that Decision Table or JRIP give the better performance with 74% with 0.25 mean absolute error. There are 10 clusters are experimented for the I-Virus dataset to find the Time taken to build model (full training data) and percentage of Incorrectly clustered instances. Among 10 clusters K-Means clustering gives the best result as only 40% instances are incorrectly clustered instances with 0.01 seconds. Therefore we can conclude that Decision Table or JRIP are the best classifiers for accuracy predictions for I-Virus influenza survivability on the basis of symptoms given in dataset among patients.

REFERENCES

[1] Kashish Ara Shakil, Shadma Anis And Mansaf Alam, “Dengue Disease Prediction Using Weka Data Mining Tool”
[2] Svetlana S. Aksenova, Machine Learning With Weka Weka Explorer Tutorial For Weka Version 3.4.3
[3] Wikipedia,http://en.m.wikipedia.org/wiki/weka (machine learning), accessed in January 2015.
[4] Waikato,http://www.cs.waikato.ac.nz/ml/weka,accessed in January 2015.
[5] Wikipedia,http://en.m.wikipedia.org/wiki/Data_set,accessed in January 2015.
[6] KirkbyR, Frank E, WEKA Explorer User Guide for version 3-4-3, November2004.
[7] Wikipedia,http://en.m.wikipedia.org/wiki/Naive_Bayes_classifier,accessed in January 2015.
[8] https://en.wikipedia.org/wiki/Influenza_A_virus



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY