Chronic Diseases Diagnosis by Classification

Ashutosh kumar sinha[1], Miss Pujasuman Tripathy[2]

[1]*M.Tech Student, CSE Department, CAPGS, BPUT, Rourkela, Odisha, India*
[2]*Faculty, CSE Department, CAPGS, BPUT, Rourkela, Odisha, India*

---***---

Abstract - *The threats to people's health from chronic diseases are always exist and increasing day by day. So it is most significant to diagnose the disease at an early stage. This paper focusses to develop a model for diagnosis of chronic disease by the help of data mining and classification technique. In this paper C4.5 decision tree has been used, which is a statistical classifier. We are taking the help of case based reasoning, embedded disease rules and rule base. Consequently all these data are processed through weka which is java based open source software. Finally we got two categories, First category possessing chronic diseases and second category not affected with chronic diseases. Apart from chronic diseases, we can apply our proposed model to non chronic diseases too. We can also use this proposed model for benefit of animals and birds.*

Key Words: **Chronic disease, Data mining, Case-based reasoning, C4.5 decision tree.**

1. INTRODUCTION

The threats to people's health from chronic diseases are increasing now a days day by day and always exist. So it is foremost important to diagnose the disease at an early stage. Diagnosis of chronic diseases is very vital as these diseases persist for long time. The leading chronic diseases include diabetes, heart disease, asthma, cancer, copd, arthritis. Early detection and diagnosis of chronic disease helps in taking preventive actions and leads to effective treatment at an initial stage. As effective and proper treatment at an initial stage has always been found to be helpful for patients. Prediction of diseases become quick and easier if data is precise and consistent. We have applied C4.5 decision tree which is statistical classifier for the diagnosis of chronic diseases. We have used classification technique which is the most important tool of data mining. Classification is basically a categorization which is a kinds of databases mined, kinds of knowledge mined. We have proposed model and have used embedded disease rules, rule base and case based reasoning and whole dataset are processed through weka. weka is a java based open source software which is inquisitive in nature. For improved decision making we use weka, as thinking machines weka does best. Weka is a collection of machine learning algorithms and can do data analysis as well. We have taken various attributes such as cholesterol, sugar level, bp, pulse which triggers a lot to chronic diseases. After processing dataset through weka we get classified data information and we get finally two basic categories, first category those are affected with chronic disease in extreme as well as in moderate manner. Second category we get are those not at all affected with chronic diseases. Our proposed model is also used for non chronic diseases and it can also be used for benefit of animals and birds.

1.1 Chronic Diseases and Types of Chronic Diseases

A disease which continue to exist for a long time that is three months or more than that. Generally it cannot be prevented by vaccination or can be cured by medication. It is generally triggered due to usage of tobacco, physical inactivity. There are various types of chronic diseases **e.g. heart disease, asthma, cancer, COPD, arthritis and diabetes**. So proper diagnosis is needed to deal with chronic diseases.

1.2 Classification and It's Various Types

Classification is basically a data mining function which allocates items in a collective form to aim categories. Classification is based upon the kinds of the databases mined, kinds of knowledge mined, kinds of techniques utilized and according to applications adapted. Proceeding with classification we are able to recognize which data belongs to which category. ID3, C4.5, Naive Bayes Algorithm, SVM, ANN are various classification techniques. In this paper C4.5 decision tree has been used which is a statistical classifier.

2. PROPOSED MODEL

Firstly we will collect chronic diseases related dataset which is preprocessed. On the other hand we go through embedded disease rules. Basically e.g. if sugar level increases and exceeds normal level then we say that person is diabetic. So increased level of sugar is termed as embedded disease rule for diabetes. Similarly increased cholesterol level yields to heart diseases and triggers heart strokes and this is embedded disease rule for heart disease and so on.

All these embedded disease rules are collected under rule base, which is storage (collection) of disease rules. As soon as we get any new case then going through embedded disease rules and rule base we predict the probability of person possessing some specific chronic diseases.

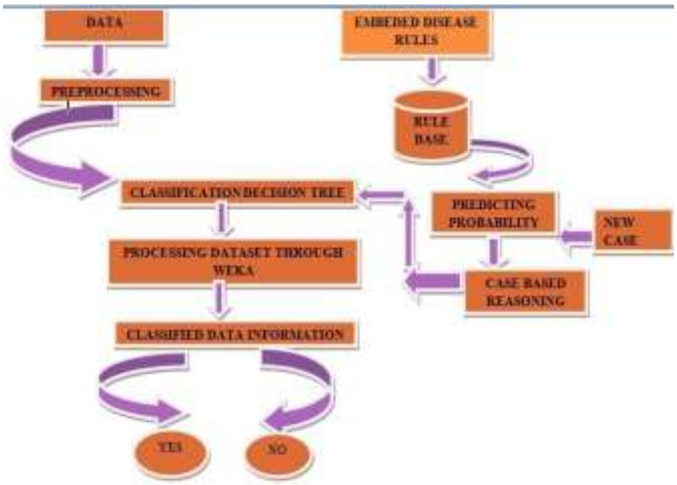

Fig -1: Proposed Model of Detection of Chronic Diseases

This probability goes through case based reasoning. CBR is the process of solving new problems based on the solutions of similar past problems. Then using C4.5 decision tree classification is done, which is statistical classifier. After that dataset are processed through weka which is java based open source software which is a collection of machine learning algorithm and can do data analysis as well and is inquisitive in nature. Consequently we get classified data information. We have taken into account cholesterol, bp, sugar level, pulse rate and so on as attributes for diagnosis of chronic diseases through weka. Finally we get two categories. First category belongs to person those are extremely, moderately possessing chronic diseases and second category belongs to people those are not at all affected with chronic diseases.

3. TOOL STUDY

3.1 Weka

Weka is basically a java based free and open source software. Weka is a collection of machine learning algorithms and as we know machine learning is a form of artificial intelligence and AI is form of computer science with software capable of self modification, programs capable of changing themselves, programs capable of improving themselves. Weka can do data analysis as well as human learning and it follows classical statistical techniques .For improved decision making we are considering Weka and it does best as thinking machines. Weka's full form is Waikato Environment for Knowledge Analysis. Weka is developed at the University of Waikato, New Zealand. In fact Weka is a flightless bird found in New Zealand with an inquisitive nature which is very curious in nature based upon that this tool has been given name as Weka.

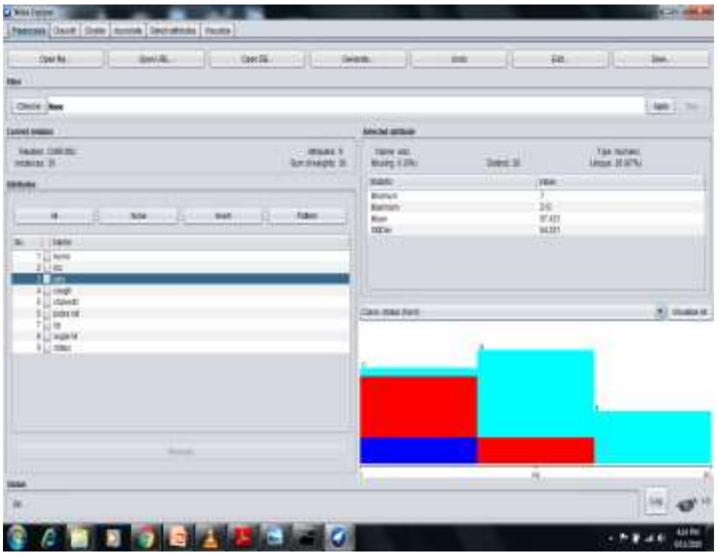

Fig -2: Weka Working Environment

The above figure shows the working environment of Weka. As we can see Weka contains tools for data pre-processing, classification, clustering, association rules and visualization. Weka consists of three windows. Left most window consists all attributes. Second window on right upper side represents label and third window on rightmost lower side represents visualization.

4. RESULT AND DISCUSSION

JUST BEFORE UPLOADING DATASETS

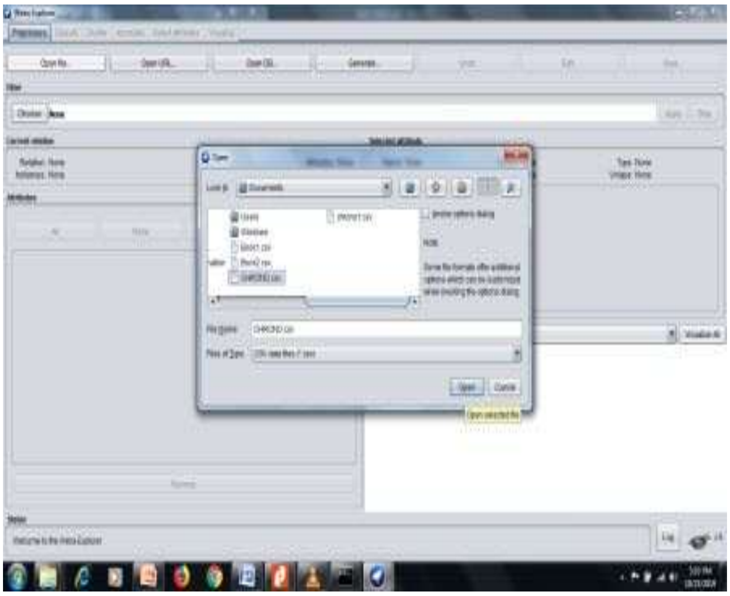

Fig -3: Screenshot Just Before Uploading Datasets

The above pic is depicting the scenario of just before uploading datasets.

JUST AFTER UPLOADING DATASETS

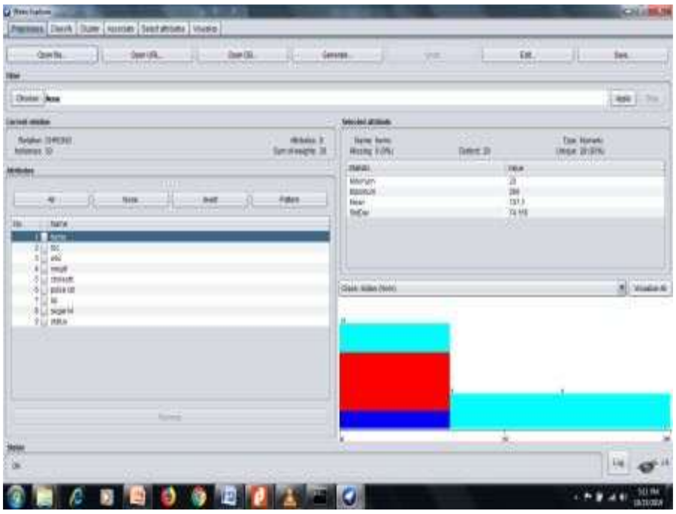

Fig -4: Screenshot Just After Uploading Datasets

The above pic is depicting the scenario of just after uploading datasets.

Thus we get the following results:

In order to diagnosis of chronic disease we have gone through various tools study. Weka is most suitable because it covers all field of data mining. Weka focuses upon preprocessing which is the first step of classification which removes redundancy. We have taken some attributes like Hemoglobin, red blood cells, white blood cells, cough, cholesterol, pulse rate, blood pressure, sugar level, status into account for the diagnosis of chronic diseases. We have those datset where we can find individual diagram of each and every attribute. As we can see here are three windows. First window shows attributes, second window represents statistical value and third window meant for visualization.

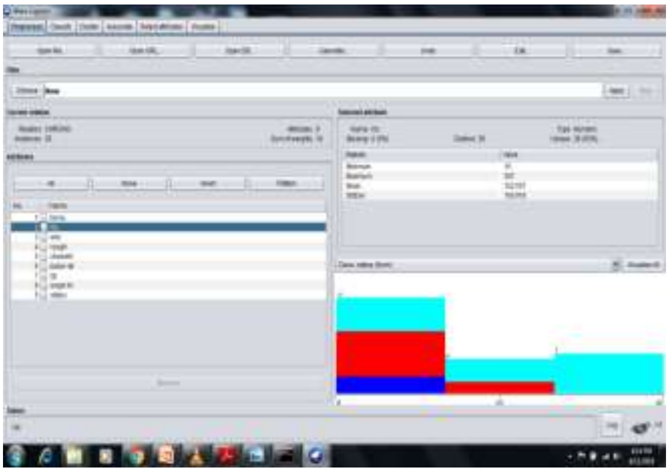

Fig -5: Visualizing Max, Moderate & Min Ranges of RBC through Weka

In the previous pic we have taken into account attribute RBC (red blood cell) which is a significant factor of chronic diseases. RBC is taken into account for Cancer.

As we can see in the figure there are different colours aqua followed by red and blue. Aqua colour represents people having maximum level of red blood cells. Red colour represents people possessing moderate level of RBC. Whereas blue colour represents people having minimum level of RBC.

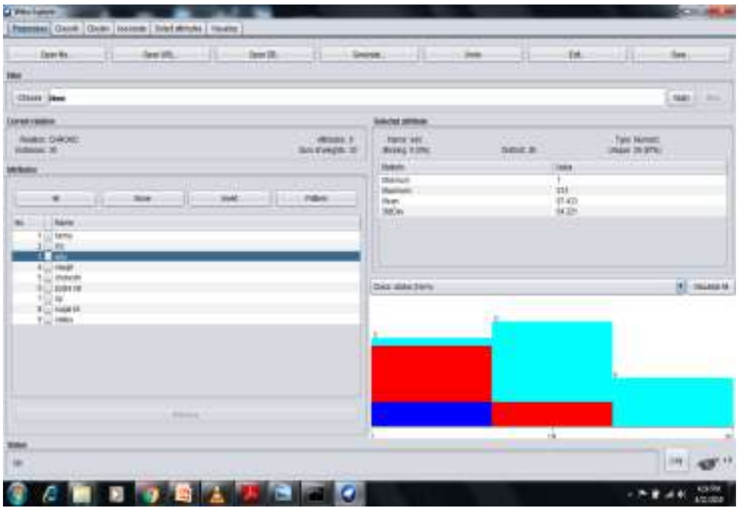

Fig -6: Visualizing Max, Moderate & Min Ranges of WBC through Weka

In the above pic we have taken into account yet another crucial factor i.e. WBC (white blood cells) for diagnosis of chronic diseases. As we know immune system is closely related with white blood cells. WBCs, also called leukocytes, are an important part of the immune system.

As we can see the figure Aqua colour represents the people possessing maximum level of wbc.Red colour represents people having moderate level of wbc.Blue colour represents people having minimum level of WBC.

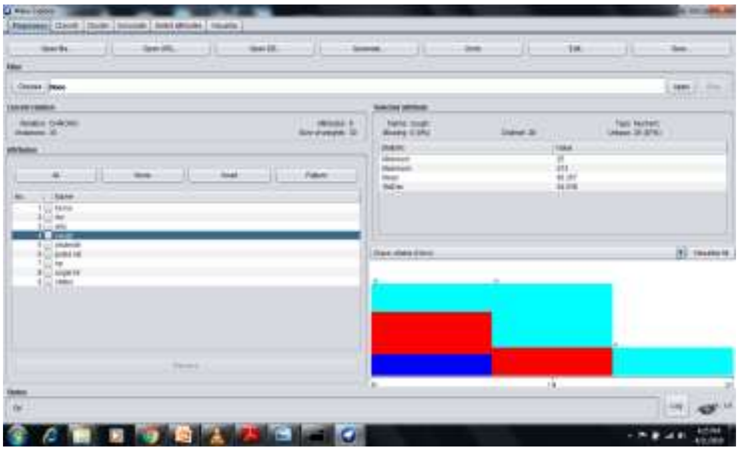

Fig -7: Visualizing Max, Moderate & Min Ranges of Cough through Weka

In the previous pic we have considered another attribute which is cough for the diagnosis of chronic diseases. Cough is the major factor of ashthma. It makes breathing difficult.

As we can in the figure, the aqua colour represents people having maximum level of cough where as Red colour represents people possessing moderate level of cough. Blue colour represents people having people having minimum level of cough.

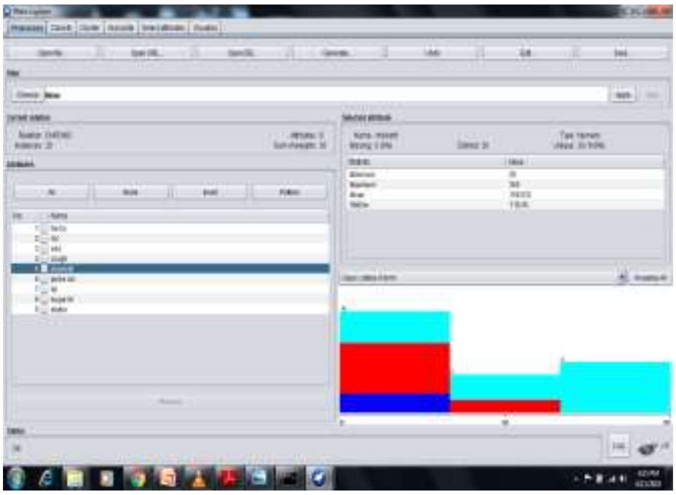

Fig -8: Visualizing Max, Moderate & Min Ranges of Cholesterol through Weka

In the above pic we have taken into account most significant attribute i.e. cholesterol if its value exceeds then it triggers heart stroke and heart diseases.

As we can see in the figure, the aqua colour represents people having maximum level of cholesterol whereas red colour represents people possessing moderate level of cholesterol .Blue colour represents people having minimum level of cholesterol.

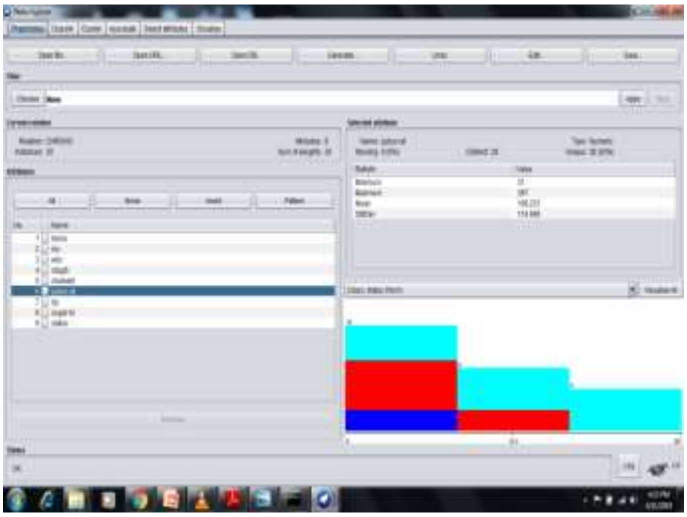

Fig -9: Visualizing Max, Moderate & Min Ranges of Pulse Rate Through Weka

In the previous pic we have taken into account pulse rate which is also most vital attribute for diagnosis of chronic diseases. If pulse rate is abnormal or very fast then it may lead to sudden cardiac arrest or heart attack. So pulse rate should be taken into account for diagnosis of chronic diseases.

As we can see in the figure there are different three distinguish colours aqua, red and blue. Aqua colour represents people having maximum level of pulse rate. Red colour represents people having moderate level of pulse rate. Where as Blue colour represents minimum level of pulse rate.

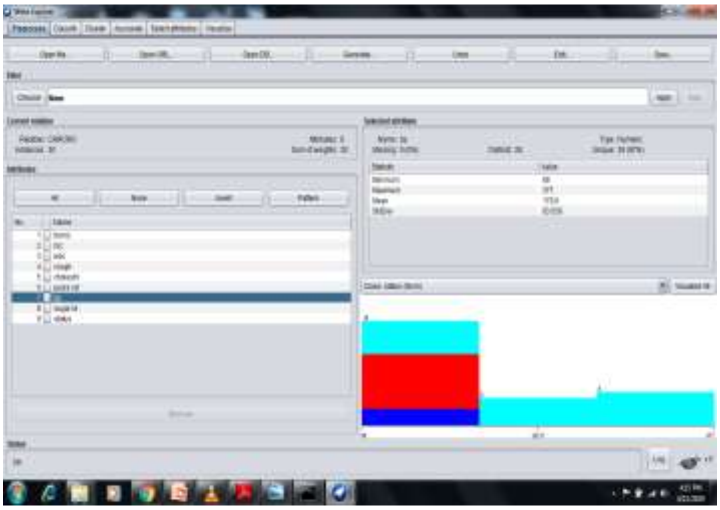

Fig -10: Visualizing Max, Moderate & Min Ranges of BP through Weka

Blood pressure triggers a lot to chronic diseases. In the above pic we have taken BP into account. Fluctuation of blood pressure may cause severe health problems. Long-term hypertension triggers heart stroke.

In the figure captured by weka it is clearly visible, Aqua colour represents people possessing maximum level of BP. Red colour represents people having moderate level of BP. Where as blue colour represents people having minimum level of BP.

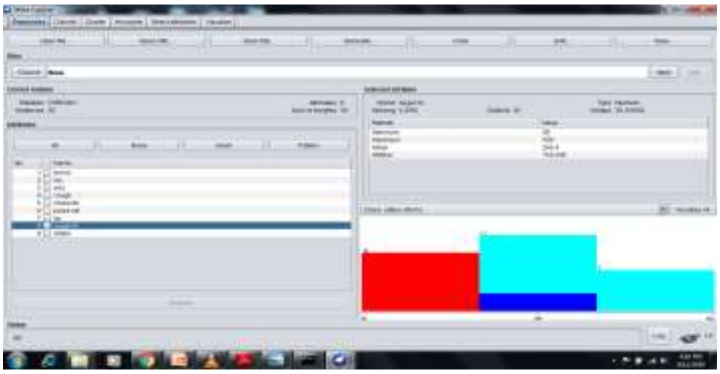

Fig -11: Visualizing Max, Moderate & Min Ranges of Sugar Level Through Weka

As we know if amount of sugar level exceeds to normal level, then we say that particular person is diabetic. So in order to diagnose diabetes, in the previous pic sugar level has been taken into account.

As we can see in the figure aqua colour represents people possessing maximum level of sugar level. Red colour represents people having moderate level of sugar level. Where as blue colour reperesents people having minimum level of sugar.

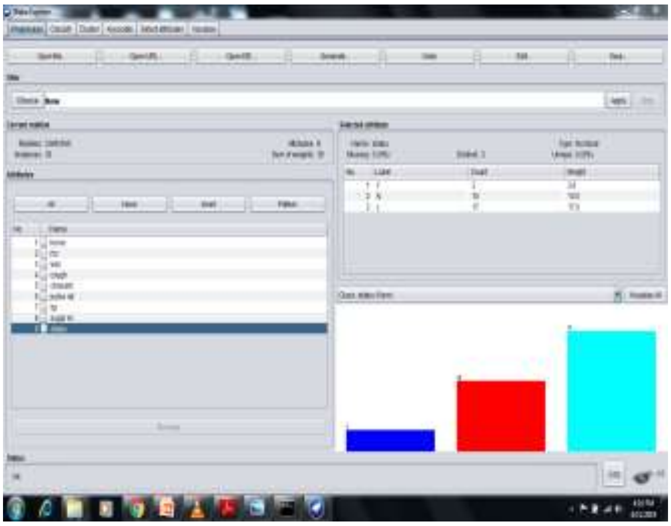

Fig -12: Visualizing Overall Status of All Attributes through Weka

Visualizing overall status of all attributes through Weka is foremost importance. Through this we can infer exactly how many peoples are affected with chronic diseases or not, taking all attributes into account. As we can see in the above pic there are three windows. First window represents all attributes, second window represents label and third window represents visualization.

As we can see clearly Aqua colour represents the exact number of peoples those are extremely affected with chronic diseases. Blue colour represents the exact number of peoples those are mildly affected with chronic diseases. Red colour represents the exact number of peoples those are not at all affected with chronic diseases.

As we have taken into account nine attributes(hemoglobin, rbc, wbc, cough, cholesterol, pulse rate, bp, sugar and status) these are directly proportional to chronic diseases. In the next pic, as we can see the single window has been splitted into nine sub windows. Each sub windows represents individual attributes of chronic diseases.

In the upcoming pic we can see, in a single window we can visualize all attributes combiningly at a time simultaneously. This is the most advantage of weka. That's the reason we prefer Weka.

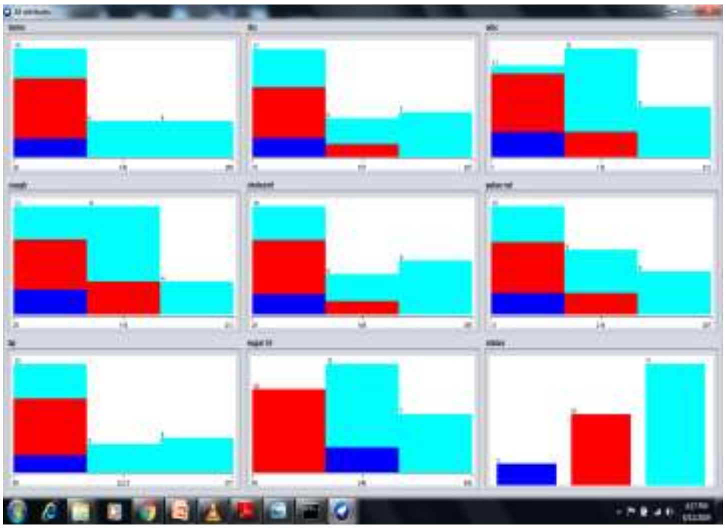

Fig -13: Overall Visualization of All Attributes Contributing to Chronic Diseases

Through Weka we can visualize two attributes and plot a graph between them. In the below pic we have taken cholesterol on x axis and sugar on y axis.

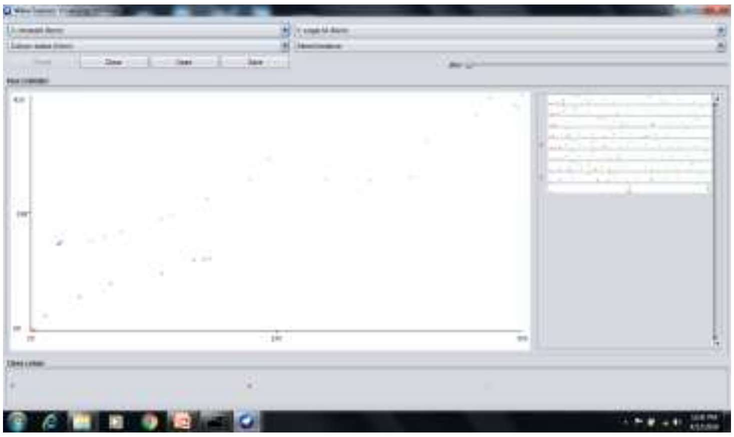

Fig -14: Visualization of Cholesterol Vs Sugar through Weka

As we can see we get three colours based crosses. Blue cross represents peoples those are mildly affected with chronic diseases while taking into account attributes cholesterol and sugar. Red crosses represents people those are not affected with chronic diseases .Green crosses represents people those are extremely affected with chronic diseases.

In the pic below we have plotted graph in between hemoglobin and cough. Hemoglobin is on x axis where as cough is on y axis.

As we can see, Blue cross represents peoples those are mildly affected with chronic diseases while taking into account attributes hemoglobin and cough. Red crosses represents people those are not affected with chronic diseases .Green crosses represents people those are extremely affected with chronic diseases.

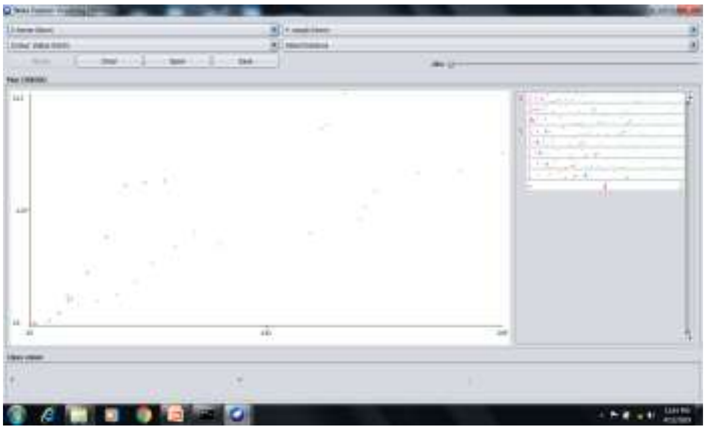

Fig -15: Visualization of Hemoglobin Vs Cough through Weka

4. CONCLUSION AND FUTURE WORK

Firstly we have determined aim/objective/problem definition. Defining problem is foremost important. Then based on objective we have done respective tool study. We have studied various tools but Weka is most compatible because it is collection of machine learning algorithms. After that we have proposed a model for chronic diseases diagnosis. We have run the data set on Weka and able to classify the parameters of chronic diseases. Going through our proposed model we can diagnose chronic disease at very early stage and can avoid early deaths.

We can apply this proposed model to other diseases like non chronic diseases also. We can also use this proposed model for benefit of animals and birds. Apart from human being our proposed model is also useful for saving lives of birds and animals. The key idea is to just focuss upon their respective relevant attributes and we have to go through all those relevant attributes and we have to process all data sets through Weka. For non chronic diseases too we have to go through their respective attributes and processing through Weka we can save many lives so that earlier death can be avoided.

REFERENCES

[1] Mu-Jung Huang a,*, Mu-Yen Chen b, Show-Chin Lee a "Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis", National Changhua University of Education, Changhua 50058, Taiwan, ROC Taiwan Expert Systems with Applications 32 (2006) 856–867.

[2] Divya Jain, Vijendra Singh "Feature selection and classification systems for chronic disease prediction: A review", The NorthCap University, Gurugram 122017 A review. Egyptian Informatics J (2018).

[3] Nini H. Jonkman, MSc, Rolf H.H. Groenwold, MD, PhD, Jaap C.A. Trappenburg, PhD, Arno W. Hoes, MD, PhD, Marieke J. Schuurmans, RN, PhD "Complex self-managemet interventions in chronic disease unravelled: a review of lessons learnt from an individual patent data meta-analysis", VU University Amsterdam Journal of Clinical Epidemiology (2017).

[4] Lars P. Thomsena,⁎, Ulla M. Weinreichb, Dan S. Karbinga, Vanja G. Helbo Jensenc,Morten Vuustd, Jens B. Frøkjærc, Stephen E. Reesa "Can computed tomography classifications of chronic obstructive pulmonary disease be identified using Bayesian networks and clinical data", Respiratory and Critical Care Group (RCARE), Denmark Prevention of COPD 10/04 (2011).

[5] Yin Kang, Jeffrey C. Fink, Rebecca Doerfler, Lina Zhou "Disease Specific Ontology of Adverse Events: Ontology extension and adaptation for Chronic Kidney Disease", Department of Medicine University of Maryland, Baltimore Computers in Biology and Medicine (2018).

[6] Lim Jia Hao a, Marhanis Salihah Omar a *, Noorlaili Tohit b "Polypharmacy and Willingness to Deprescribe Among Elderly with Chronic Diseases", University Kebangsaan Malaysia.

[7] Carlton Dampier, MD, Tonya M. Palermo, PhD, Deepika S. Darbari, MD, KathrynHassell, MD, Wally Smith, MD, William Zempsky, MD "AAPT Diagnostic Criteria for Chronic Sickle Cell Disease Pain", Seattle Children's Research Institute USA Journal of Pain(2017),

[8] Jin Hee Kim a, Ki Young Son a, Dong Wook Shin a, Sang Hyuk Kim a, Jae Won Yun b,c, Jung Hyun Shin a,Mi So Kang a, Eui Heon Chung a, Kyoung Hun Yoo a, Jae Moon Yun a, "Network analysis of human diseases using Korean nationwide claims data", Samsung Medical Center, Seoul, Republic of Korea.

[9] Liqin Wanga,b, Peter J. Haug a,b, Guilherme Del Fiol a "Using classification models for the generation of disease-specific medications from biomedical literature and clinical data repository", Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Salt Lake City, UT 84108, USA.

[10] S.B.Akben "Early Stage Chronic Kidney Disease Diagnosis by Applying Data Mining Methods to Urinalysis, Blood Analysis and Disease History", Osmaniye Korkut Ata University, Osmaniye, Turkey.

[11] Saif Nalband Aditya Sundar A. Amalin Prince Anita Agarwal "Feature Selection and Classi_cation Methodology For The Detection of Knee-Joint Disorders", Birla Institute of Technology and Science, Pilani.

[12] Dr Saravana kumar N, Eswari, Sampath & Lavanya, "Predictive Methodology for Diabetic Data Analysis in Big Data", Sri Krishna College of Engineering & Techechnology, Coimbatore.

[13] Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqui, "Application of data mining: Diabetes health care in young and old patients" College of Computer Engineering and Sciences, Salman bin Abdulaziz University, Saudi Arabi, jksuci.2012.10.003.

BIOGRAPHIES

Ashutosh Kumar Sinha has completed M.Tech in CSE specialization in August 2019 at CAPGS, BPUT, ROURKELA, ODISHA

Miss Pujasuman Tripathy is Faculty in Computer Science Dept. at CAPGS, BPUT, ROURKELA, ODISHA.