National Conference on Engineering Innovations and Solutions (NCEIS – 2018)
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 4 | Issue 6 | ISSN : 2456-3307

Prediction Of Diseases Using Haematology Records

Vivek S Bharadwaj, Vidyashree N, Sunayana H M, Praveen Mane U, Mr. Prashanth M V
Final Year Students, Department of ISE, Vidyavardhaka College of Engineering, Mysuru, Karnataka , India

ABSTRACT

Health information system is a system used for decision making such as capturing, storing, transmitting information which is related to the health of particular individual or the activities that are associated to health steam. The health information system collects data from relevant health sector that analyze, check the quality, relevance and timeliness, and then converts data into information for decision-making.. A good health information system gathers all the data to make sure that users of health information have access to data. Sound decision-making at all levels of a health system requires reliable health statistics that are disaggregated by sex, age and socioeconomic characteristics. Decisions along with proof contribute to more efficient resource allocation and information about the quality and effectiveness of services can contribute to better outcomes in health information system.

Keywords: Hematology records, Clustering algorithms, Prediction, Accuracy

I. INTRODUCTION

Haematology disease is the science or study that makes the blood the center of concern. This science studies the components of blood, blood forming organs and diseases caused by blood.

Analysis of haemogram blood test and Liver Function Test (LFT) samples can be used to predict the diseases using data mining techniques such as K-means, Fuzzy C means. Among data mining techniques, clustering technique is considered as one of the efficient DM technique, which groups the data items based on similarities and dissimilarities. Several similarity and dissimilarity measures are applied to find the relationships and patterns which exist in data items. This paper focuses to extract hidden rules and relationships between diseases from a real-world Healthcare Information System. The information regarding the presence or absence of a disease is only used. The parameters used for the disease prediction are cholesterol level, blood sugar, blood pressure and so on. The method also predicts disease based on other diseases that a patient already has. The advantage of this approach is that it can be applied to predict any disease rather than a specific disease by using the parameters.

Parameters considered for Complete Blood Count (CBC) and Liver Function Test (LFT)-

Complete Blood Count Parameters	Liver Function Test Parameters
-Haemoglobin (Hb) -Red Blood Cells (RBC) -White Blood Cells (WBC) -Neutrophils or Polymorphs -Eosinophil -Basophils -Lymphocytes -Monocytes -Platelets	-Serum Bilirubin -Alkaline Phosphate (ALP) -Serum glutamic oxaloacetic transaminase (SGOT) -Serum glutamic pyruvic transaminase (SGPT) -Total Proteins -Albumin -Globulin

-Mean corpuscular volume (MVC) -Mean corpuscular haemoglobin (MCH) -Mean corpuscular haemoglobin concentration (MCHC)	

Parameters considered for Complete Blood Count (CBC) are:

-**Haemoglobin percentage**:

Males -Normal: 14-18 gm%

Average: 15.5 gm%

Females- Normal: 12-15.5 gm%

Average: 14 gm%

Probable disease-Less than normal- Anaemia

More than normal- Polycythaemia vera

-**Red Blood Cells**:

Males- Normal: 5.6 million/ML

Average: 5.5 million/ML

Females- Normal: 4.5-5.5 million/ML

Average: 4.8 million/ML

Probable disease-Less than normal- Anaemia

More than normal- Polycythaemia vera

-**White Blood Cells:**

Normal: 4000-11000/ML

Probable diseases-Less than normal-Thyroid, Viral infection, Protozoan infection, Bone marrow depression.

-**Neutrophils or Polymorphs:**

Percentage: 50-70%

Absolute count: 3000-6000/ML

Probable diseases-Less than normal-Typhoid fever, Viral infection

More than normal- Bacterial infection, Myocardial infection

-**Eosinophil:**

Percentage: 1-4%

Absolute count: 150-300/ML

Probable diseases-More than normal- Bronchial asthma, Warm infestation

-**Basophils:**

Percentage: <1%

Absolute count: 10-100/ML

Probable diseases-More than normal- Chickenpox, Smallpox, Tuberculosis, Influenza

-**Lymphocytes:**

Percentage: 20-40%

Absolute count: 1500-2700/ML

Probable diseases-Less than normal- AIDS (HIV)

More than normal- Tuberculosis, Lymphatic leukaemia, Viral infection

-**Monocytes:**

Percentage: 2-8%

Absolute count: 300-600/ML

Probable diseases-Less than normal-Hypo-plastic bone marrow

More than normal- Tuberculosis, Syphilis, Leukaemia

-**Platelets:**

Normal: 1.5-4 lakhs/ML

Probable diseases-Less than normal- Bone marrow depression, Dengue fever

More than normal- Trauma, Surgery, Injury, Splenectomy

-**Mean corpuscular volume (MVC):**

Normal: 78-84 FL

Probable diseases-Less than normal- Iron deficiency anaemia

More than normal- Megalablastic anaemia

-**Mean corpuscular haemoglobin (MCH):**

Normal: 28-32pg

Probable diseases-Less than normal- Iron deficiency anaemia, Thalassemia

-**Mean corpuscular haemoglobin concentration (MCHC):**

Normal: 32-35%

Probable diseases-Less than normal- Iron deficiency anaemia, Thalassemia

More than normal- Megalablastic anaemia

Parameters considered for LFT-

-**Serum Bilirubin:**

Total Bilirubin: 0.2-1 mg/dl

Direct Bilirubin: 0.1-0.3 mg/dl

Probable diseases-More than normal- Hepatitis, Haemolytic disease, Gilbert's disease, Biliary obstruction, Liver failure, Cirrhosis

-**Alkaline Phosphate (ALP):**

Normal: 33-36 U/L

Probable diseases-More than normal- Hepatitis, Bone diseases

-**Serum glutamic oxaloacetic transaminase (SGOT):**

Normal: 12-38 U/L

Probable diseases-More than normal- Hepatitis, Liver failure, Cirrhosis, Myocardial infarction

-**Serum glutamic pyruvic transaminase (SGPT):**

Normal: 7-41 U/L

Probable diseases-More than normal- Hepatitis, Cirrhosis

-**Total Proteins:**

Normal: 6-8 gm%

Probable diseases-Less than normal- Hepatitis, Cirrhosis, Liver failure

-**Albumin:**

Normal: 3.5-5.5 gm%

Probable diseases-Less than normal- Hepatitis, Cirrhosis, Liver failure

-**Globulin:**

Normal: 2-3.5 gm%

Probable diseases-More than normal- Hepatitis, Cirrhosis Performances of the clustering algorithm are analyzed by factors like accuracy, execution time and error rate.

- Time factor describes the amount of time needed for predicting the disease.
- Accuracy is verifying whether all the data items are grouped or clustered correctly to their respective groups.
- Error rate is identifying the percentage of data items which has been placed incorrectly in the set of clusters.

All these factors are compared using the algorithms like K-means, Fuzzy C means, Random Forest Algorithm, Naïve Bayes Algorithm and Support Vector Machine Algorithm

K-means clustering algorithm-K-means clustering is a type of unsupervised learning, which is used when you have data without defined categories or groups. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K . The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are: The centroids of the K clusters, which can be used to label new data. Labels for the training data (each data point is assigned to a single cluster.

Fuzzy C means-This algorithm works by assigning membership to each data point corresponding to each cluster centre on the basis of distance between the cluster centre and the data point. More the data is near to the cluster centre more is its membership towards the particular cluster centre. Clearly, summation of membership of each data point should be equal to one. It is frequently used in pattern recognition.

Random Forest Algorithm- Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because it's simplicity and the fact that it can be used for both classification and regression tasks.

Naïve Bayes Algorithm- The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

Support Vector Machine- Support Vector Machine (SVM) is primarily a classifier method that performs classification tasks by constructing hyperplanes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables.

II. CONCLUSION

A large volume of data is generated in the healthcare sector and doctor has to come to a direct contact the patient. Instead the data mining tools and classifying algorithms can be used to predict the diseases. Further the tests considered in this paper are LFT and CBC. The expected success rate of our proposed model is in the range of 65-70%. The same procedure can be applied to all the blood tests and enhance the probability of prediction. The future scope of the model includes usage of model in remote places and in places where health care facility is not accessible.

III. REFERENCES

[1] An Efficient Clustering Algorithm for Predicting Diseases from Hemogram Blood Test Samples S. Vijayarani and S. Sudha Indian Journal of Science and Technology, Vol 8(17), DOI: 10.17485/ijst/2015/v8i17/52123, August 2015

[2] Application of machine learning for hematological diagnosis Gregor Gunčar1, Matjaž Kukar1, Mateja Notar1, Miran Brvar2,3, Peter Černelč4,Manca Notar1, Marko Notar1

[3] Predicting Disease By Using Data Mining Based on Healthcare Information System Feixiang Huang, Shengyong Wang, and Chien-Chung Chan University of Akron

[4] Blood Diseases Detection Using Data Mining Techniques,2017 8th International Conference on Information Technology (ICIT), Asem H. Shurrab and Ashraf Y. A. Maghari T. N. Hubbard Faculty of Information Technology Islamic University of Gaza Gaza, Palestine.