DISEASE PREDICTION THROUGH SYNDROMES USING K-MEANS ALGORITHM

Md Aliul Islam Abir
ID: 142-15-137

This report is presented as partial fulfillment of the requirement for a bachelor's degree in Computer Science and Engineering.

Supervised By

Mr. Ohidujjaman Tuhin
Senior lecturer
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Md. Reduanul Haque
Senior lecturer
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
06th July 2019

APPROVAL

The project, titled "**Disease prediction through syndrome using the K-Means clustering algorithm**" submitted to Daffodil International University by the Dept. of Computer Science and Engineering by Md. Aliul Islam Abir, has been accepted as satisfactory for partial fulfillment of the requirements for the degree of Computer Science and Engineering. The approved presentation was held on July 06th, 2019.

BOARD OF EXAMINERS

_____

Dr. Syed Akhter Hossain, Chairman, Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Dr. S M Aminul Haque, Internal Examiner and Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Md. Reduanul Haque, Internal Examiner and Senior lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

_____

Dr. Mohammad Shorif Uddin, External Examiner and Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

I hereby declare that this project has been done under the supervision of **Md. Ohidujjaman Tuhin, Senior lecturer, Department of Computer Science and Engineering (CSE), Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for any award of any degree or diploma.

Supervised by:

_____

Mr. Ohidujjaman Tuhin
Senior lecturer
Department of CSE
Daffodil International University

Co-supervised by:

_____

Md. Reduanul Haque
Senior lecturer
Department of CSE
Daffodil International University

Submitted by:

_____

(Md Aliul Islam Abir)
ID: 142-15-137
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, I'm expressing my sincere thanks and gratitude to the Almighty for His divine blessing which gives me enough strength to complete my final year project.

I'm grateful to have Mr. Ohidujjaman Tuhin and Md. Reduanul Haque Lecturers, Department of CSE, Daffodil International University, Dhaka in this project with me. Deep Knowledge & keen interest of our supervisor and co-supervisor in the field of "Data Mining" encouraged me to continue and complete this project successfully. Their patience and energetic supervision at all stages have made it possible to complete this project.

I am grateful to Daffodil International University, Permanent Campus, for providing a natural environment that is conducive to research and quality work.

I would like to thank my classmates at Daffodil International University, who gave me valuable suggestions during this project work.

Finally, I must acknowledge with due respect the constant support and patience of my parents, and all the individuals who are directly or indirectly involved in the successful completion of this project work.

ABSTRACT

Machine learning offers a principled approach for developing sophisticated and automatic algorithms to analyze high-dimensional and multimodal biomedical data. This study focuses on using machine learning algorithms to improve detection and diagnosis of human disease. Human disease evaluation was never been easy and still a complicated process and requires a high level of expertise. Several decision support system demonstrated promising diagnostic performances in formal evaluations but only a few have been formally evaluated in clinical environments. Stand-alone decision support systems depend heavily on a vast amount of data. This study describes a research work aiming to find out how much efficient k-means can be to build an expert system to detect human disease by evaluating symptoms data to improve the quality of health evaluation. Healthcare professionals and practitioners can also use this to corroborate diagnosis. This proposed system also evaluates its performance and effectiveness and exhibits satisfactory result.

Keywords:

Data Mining, Clustering, K-means, Symptom, Diseases

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLE

TABLES **PAGE NO**

CHAPTER 1

Introduction

This chapter is about to discuss some basic criteria to clarify the overview of our work. Such as Data mining with issues, Clustering, K-means and a little bit of discussion about data and disease analysis.

1.1 Data Mining

Data mining is a technique of analyzing and finding patterns of information from random and disoriented data and transform this useful information for future use [1]. Data Mining can be considered as an interdisciplinary field involving concepts from Machine Learning, Database Technology, Statistics, Mathematics, Clustering and Visualization among others [2]. Data mining is a multi-stage process as [3] shown in Fig.1.

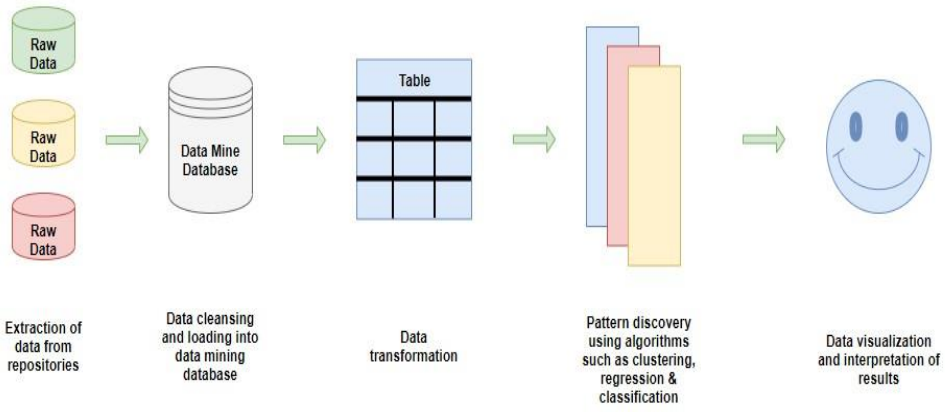

Figure 1. 1: Data mining

1.1.1 Data Mining Model

Data mining model is mainly divided in two parts which are descriptive and predictive. Four tasks of descriptive model are clustering, association rule, sequence discovery and summarization. On the other hand four tasks of predictive model are classification, time series analysis, prediction and regression.

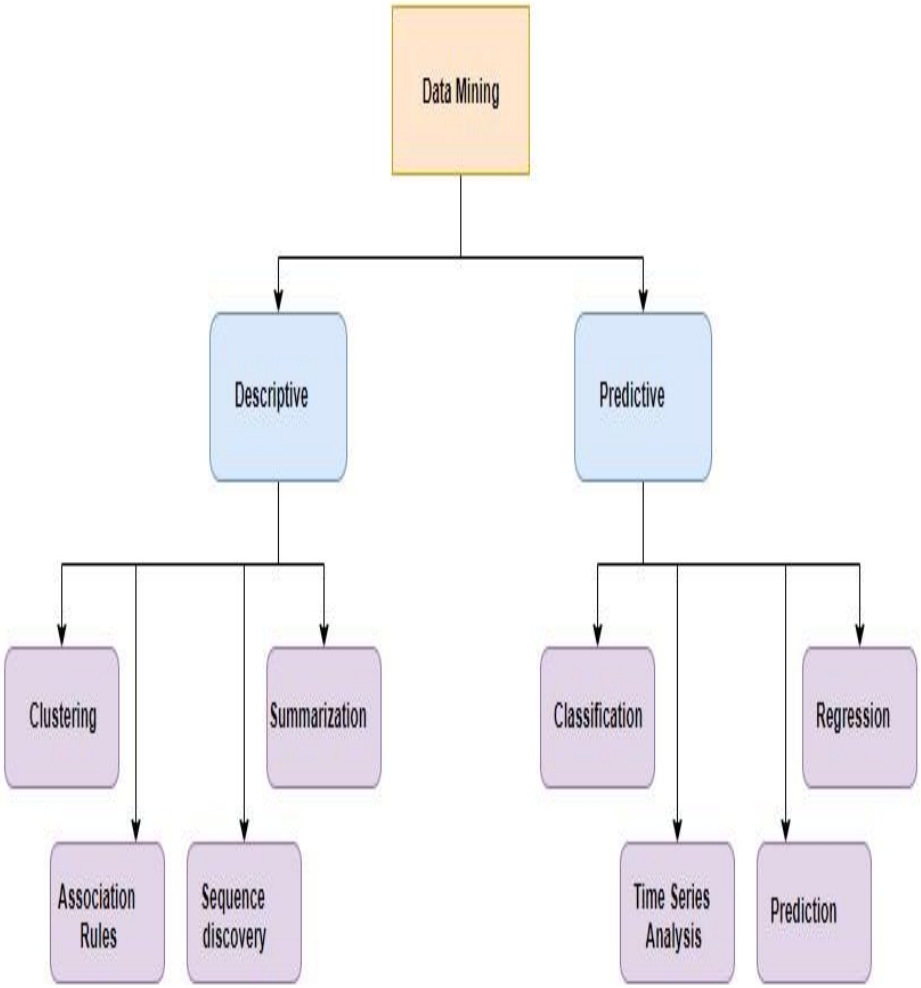

Figure 1. 2: Data mining model

1.2 Clustering

Data mining can be done by two learning approaches - supervised and uneducated education. Clustering Data Mining is an unsupervised learning application. [4] A set group of clustering objects that have the same clusters of the objects different from each other and different from the objects in different clusters. There are several fields in clustering techniques that include several fields including synthetic intelligence, pattern recognition, biological information science, division and machine learning.

Clustering is helpful in many investigative patterns-analysis, grouping, decision making and information-related situations, information recovery, image classification. However, those problems can reach a little earlier record (e.g. statistical model) and guess as much as possible about the determinant's statistics. [5] The term "clustering" is used to describe the techniques of grouping similar data altogether. Typical clustering mechanism includes the following steps [Jain and Dubes 1988]:

- Feature extraction and selection including identifying pattern
- Measuring pattern proximity
- Feature grouping
- Feature abstraction (if necessary)
- Output assessment (if necessary)

1.3 K-Means Clustering

K-Means clustering is an unsupervised learning mechanism, and used when you have data without labeled category (for example, statistics without described labels or categories). The purpose of clustering is to discover organizations within the information, to represent diverse groups of variables. The algorithm which can be fully assigned to each information factor of the organization can be determined on the basis of which they have provided. Data points are grouped together on the basis of characteristic similarity. The consequences of this algorithm are:

- Centroids of the clusters can be used to label new unknown data
- Each statistical data point can be assigned to a single cluster

Clustering lets you identify and analyze groups of data. Each series of cluster functions is a series of values that outline the next groups. Each cluster representing the crew can be tested weight centroid to quantify the quality of any form.

In clustering, measuring the distance between each point can be done in several ways such as Euclidean, Manhattan, Minkowski, etc. Euclidean distance is defined as

$$d(i,j) = \sum_{i=0}^{n} \sqrt{(x_i - y_i)^2}$$

Where, x_i and y_i are two individual data point.

Manhattan distance is defined as

$$d(i, j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots\ldots + |x_{in} - x_{jn}|$$

Minkowski distance is a simplified form of Euclidean and Manhattan distance.
And it's defined as

$$d(i, j) = (|x_{i1} - x_{j1}|p + |x_{i2} - x_{j2}|p + \ldots\ldots + |x_{in} - x_{jn}|p)1/p$$

Where, p is an integer.

The K-means algorithm consists of two separate phases.

- Select the number of centroid (k) randomly, where the value k can be determined in advance. Though there are some methods available to determine the optimal number of cluster such as Elbow method. But it depends on the purpose of the study.

- Each data point grouped together with the nearest centroid.

We have used 'Euclidean distance' in our study to measure distances between data points and cluster centroid.

Input:

K: Number of clusters.

D: Dataset containing n number of datapoints.

Output:

Set of cluster.

Method:

- ✓ Choose k number of data points from as initial random cluster centroid.
- ✓ Group similar data points as clusters on the basis of measured distances with centroids.
- ✓ Find new centroid by calculating mean of the data points in a cluster.
- ✓ Again assign each data point to the cluster based on the distance from the new cluster centroid.
- ✓ Repeat until the data is converged.

Advantages of k-means algorithm:

- Simplicity
- Its speed which allows running large datasets.
- Converge much Faster than hierarchical clustering if number of cluster is reasonable.
- It produces tighter clusters compared to hierarchical clustering.

1.4 Clustering in Disease Analysis

Data clustering is already a tested way to optimize massive data in each and every vicinity of records evaluation or data mining. In the point of view of clinical facts analyzing we must say that there's a numerous work performed before to work as a standalone solution

in the medical vicinity but none of them used to be able to take care of the entire process. But in individual area of medical science, all of them achieved sufficient to show that it's feasible to optimize information in a required way to get the job finished and data clustering is one of the fine approaches to do that. We are going to use clustering in our location of work to optimize information and the solution to get the satisfactory of it. We will use K-means clustering with the classifier.

Two of the most imperative and well-generalized issues of medical facts are its new evolved feature and concept-drift. Since a clinical records is a quick and continuous event, it is assumed to have infinite length. Therefore, it is tough to store and use all the historical facts for training. [8], [7] the most find out choice is an incremental learning technique. Several incremental rookies have been proposed to address this hassle. In addition, concept-drift takes place in the movement when the underlying concepts of the move changes over time. [9-10], [7] A range of methods have additionally been proposed in the literature for addressing concept-drift in statistics clustering. Concept-evolution takes place when new instructions evolve in the data. Cluster oriented ensemble classifier used to minimize function contrast hassle in clinical disease facts classification.

1.5 Research objective

- Disease detection using symptoms
- Initiative of primary treatment
- Assist to preliminary tests
- Reduce medical diagnosis cost
- Performance of k-Means in medical domain

CHAPTER 2

Literature Review

In this area we reviewed the preceding associated studies in the discipline of clinical science for the prediction of critical disorder based on data mining technique.

2.1 Related Works

[17]In this paper coronary heart disease prediction is executed by the use of data mining techniques. A comparative approach was taken between decision trees, neural network, and naïve Bayes. This study was carried out on the .net platform.

[18]This study taken an approach to optimize the complexity of extracting the medical data and proposed an application of medical record mining.

[20]This study developed a machine learning model using Naïve Bayes that helps nurses and medical students to deal with the patients.

[21]In this paper a machine learning model was developed using K-Means that can successfully predict heart diseases.

[22]This study explained the useful attributes of big data analysis in healthcare domain and focuses on cost-effective healthcare. An effective statistical data mining was conducted based on clinical informatics including transport cost.

[23]G. Santhosh Kumar, Lakshmi K.S proposed a new model of association rules using medical transcripts. The extracted guidelines described several relational facts between many diseases including signs of a particular disease, medical drugs used in treatment and the relation between ages with diseases.

[24]In this study a short guideline was proposed along with the principle aims of EMRs then the existing situation of EMRs is reviewed.

[25]This paper introduced Predictive information mining for scientific diagnosis. The test results are performed to compare the performance of mining methods, and it reveals that the decision tree outperform and the Bayesian classification of the category contain the same accuracy as the decision tree. However, neural networks performed better than any other mining techniques based on performance report.

[26]In this paper, the author proposed a model using K-Means clustering for predicting cardiovascular diseases with an actual and artificial dataset. Cluster analysis technique that monitors the ambition to run clusters above and closest to each crew or cluster. All clusters start from the allocation and randomization. In addition, the distance between elements and related cluster syndrome data was reduced by the square sum reduction. Research results indicate that clustering unification provides the best possible accuracy and firmly committed results.

[27] In this study, the authors proposed a search engine to resolve search queries for papers in a less amount of time. Authors developed a search engine based on quick reading algorithms that spends less time and provide quality results. An enhanced K-Means clustering approach was taken to make the algorithm more efficient and effective. Get the right cluster of documents with less complexity.

[28] In this study, the authors proposed a model using K-Means clustering to predict myocardial infarction (MI). This model extracts hidden information from the historical data of heart diseases. The functionality of the model is improved by K-Means. This method is one of the highest standard for predicting heart diseases.

[29]In this study, the authors proposed a model to examine the behavior of users on the internet based on a network log. K-means clustering algorithm is used to build the model. K-means is used primarily for clustering no. of visitors that are segmented based on the usage of low, medium, and the high volume of data then processed further. The result shows promising aspects of identifying internet activity based on normal searches, social media activity, news, podcast, etc. It also helps to identify the amount of traffic on a particular website.

[31]In this paper, the authors used K-Means clustering to identify several diseases such as heart diseases, liver cirrhosis, diabetes, cancer, etc. Then a comparative analysis was done with cautions.

CHAPTER 3

Overview of Diseases

Whenever a patient feels any kind of illness at the same time he or her body shows some symptoms related to the disease. When a patient meets to the doctor, normally doctor asks the patient what happened with him or her and try to examine primary physical condition. In this process a symptom can be caused for several diseases, so making a decision on the basis of primary external examine of a patient is so hard for a doctor that's why doctor normally gauss more than three option for a certain patient and suggest the patient for doing lots of pathological examination. In this process doesn't cost and time effective for a patient but it is more effective for the doctor for making his decision about the patient disease.

In this chapter we are going to talk about overall process how a doctor decide certain disease on the basis of different symptoms. For this research we collected 61 symptoms and there certain twenty diseases.

Figure 3. 1: A doctor examine her child patient

3. 1 Anemia

Anemia is the most common blood disorder disease in the world. Anemia is a health situation when the present doesn't have enough healthy red blood cell to transfer adequate oxygen to the patient body's tissue. Having anemia in a patient body may show five kinds of main symptoms like fatigue, weakness, pale skin, shortness of breath and tachycardia.

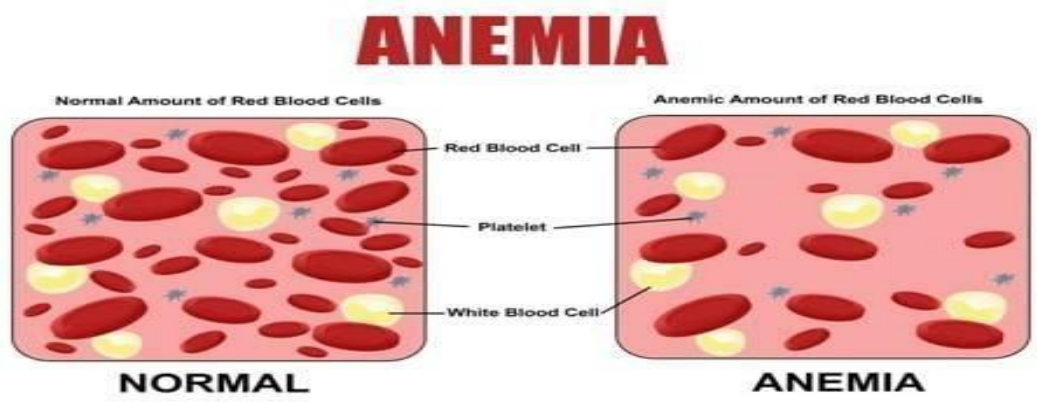

Figure 3. 2: Red blood cells comparison normal and anemia diseased.

3.2 Angina

Angina is a common disease and it is related to heart disease. Basically, angina is one kind of chest pain and it is created for low blood flow to the heart. After happening angina there are four main symptoms can be shown. Symptoms are chest pain, pain in the arms, neck, jaw, shoulder or back, shortness of breath and dizziness.

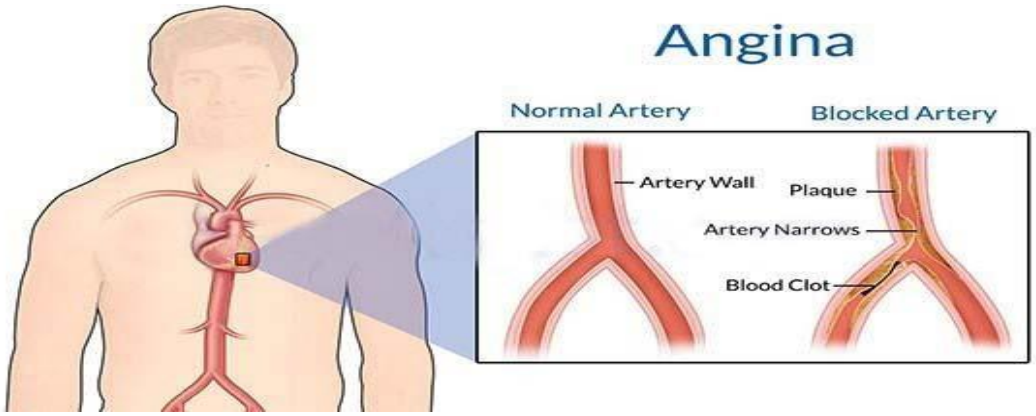

Figure 3. 3: Artery comparison normal and angina diseased.

3.3 Asthma

Asthma is one sort of chronic lung disease, and it happened for inflamed and swollen in the bronchial tube. After happening asthma, the patient may experience three main symptoms like shortness of breath, chest tightness and wheeze.

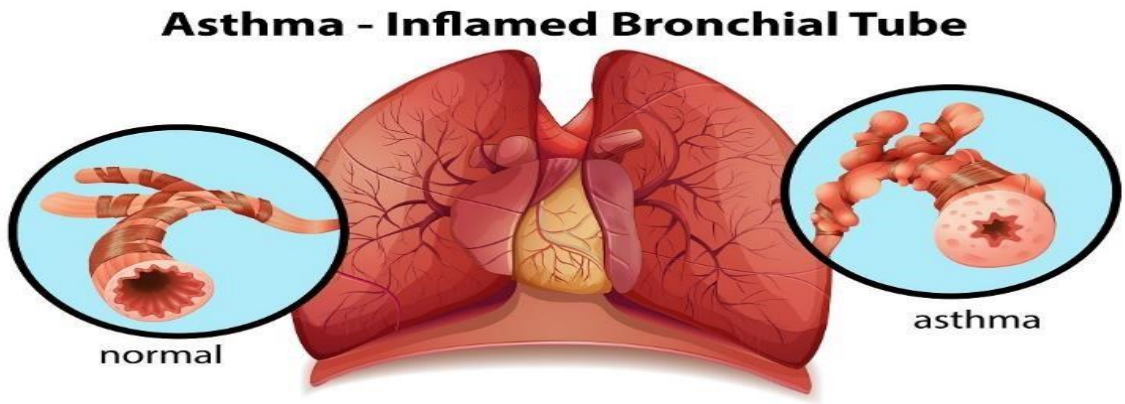

Figure 3. 4: Bronchial Tube comparison normal and asthma diseased.

3.4 Bacillary Dysentery

Dysentery has two types, and bacillary dysentery is one of them. This happens from an intestinal infection in the human gut and the intestinal infection caused by a group of bacteria called Shigella bacteria, so bacillary dysentery also called shigellosis. Having this disease shows some main symptoms like loose motion with blood and mucus, abdominal cramping and fever.

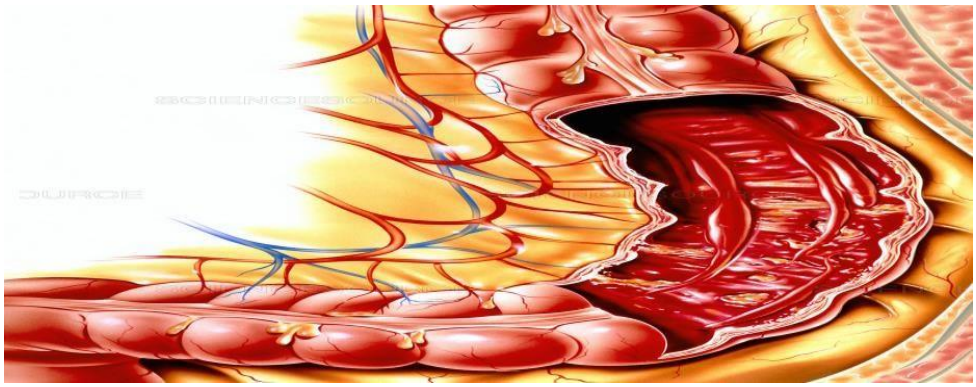

Figure 3. 5: Bacillary Dysentery disease.

3.5 Bronchiolitis

Bronchiolitis is an illness which most of the patient are kids. This is an injury of the respiratory tract and it's an infection of bronchioles that make the normal airways to the small airways and it caused by a viral infection. Most common virus infection is the respiratory virus. The symptoms of bronchiolitis are runny nose, cough, rhonchi, and fever.

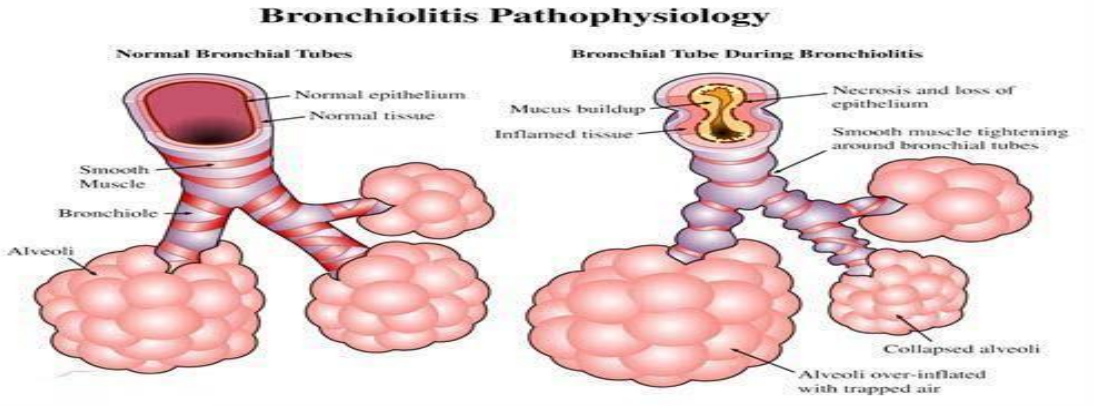

Figure 3. 6: Bronchial Tube comparison normal and bronchiolitis diseased.

3.6 Chickenpox

Chickenpox is a virus caused disease and it's a childhood common illness. The varicella zoster virus (VZV) is the main responsible for this disease that's why another name of chickenpox is varicella. The main symptoms of chickenpox are blisters, papules, crust, fever, and malaise.

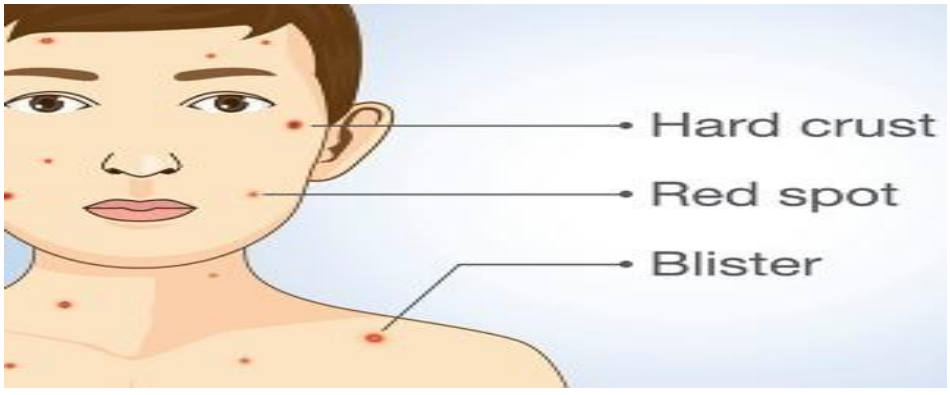

Figure 3. 7: Chickenpox disease.

3.7 Dengue Fever

Dengue fever is a viral illness that happens from the infection of dengue virus. This virus is transmitted by certain female mosquitoes. Especially the species Aedes aegypti contain this virus. There are six main symptoms can be shown on the patient like fever, headache, muscle bone and joint pain, rash, abdominal pain, and bleeding gums or nose.

Figure 3. 8: Aedes aegypti mosquito.

3.8 Diabetes Mellitus

Diabetes mellitus is a state of impaired metabolism of carbohydrate, protein, and fat happen by either leak of insulin secretion or decreased the sensitivity of the tissue to insulin. It has six main symptoms like polydipsia, proteinuria, polyuria, and hyperglycemia. Normally it has two type.

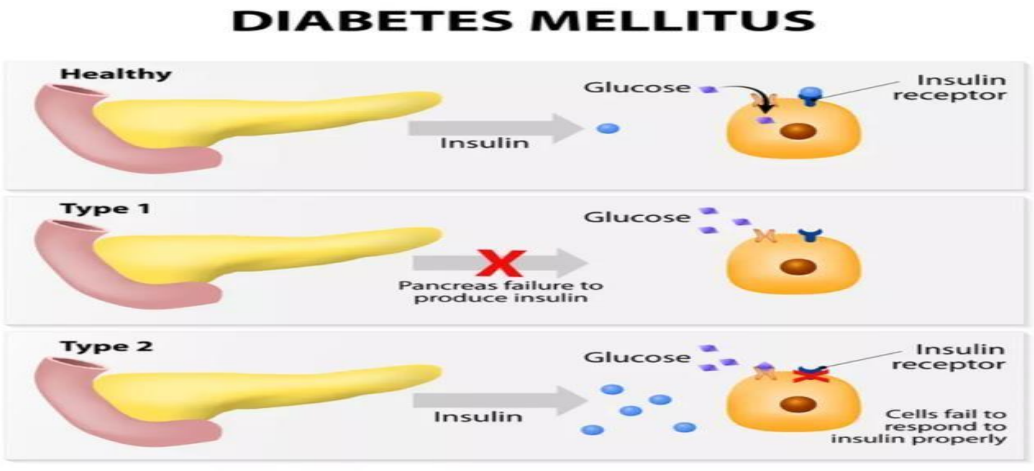

Figure 3. 9: Diabetes mellitus.

3.9 Diarrhea

Diarrhea is one of the most common diseases across the world. When a patient has to be the passage of loose stool and watery stool three or more times in a day is known as Diarrhea. Loose motion, dehydration, abdominal cramping, abdominal pain, and fever are the main symptoms of diarrhea.

Figure 3. 10: Diarrhea.

3.10 Jaundice

Normally jaundice is a symptom of underlying disorders and it is a common condition of the newborn child. When jaundice has on a body, it shows yellow appearance in the skin, sclerae and mucus membrane. It happens from a high amount bilirubin concentration in the body fluids. The symptom of jaundice is the yellow coloration of urine, the yellow colored sclera of eyes, rice watery stool, nausea, and vomiting.

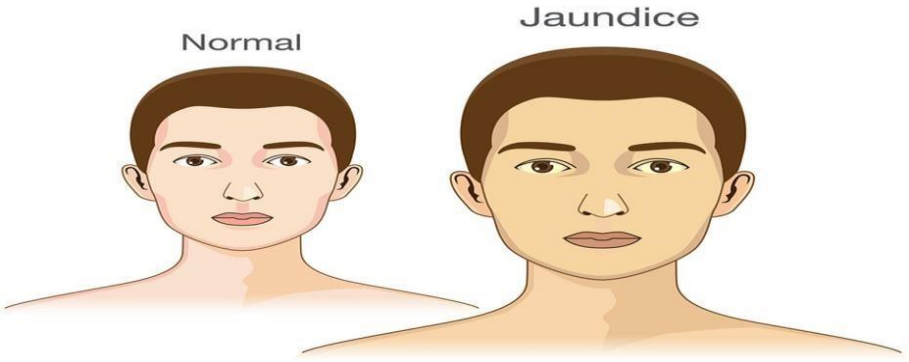

Figure 3. 11: Comparison of normal and jaundice diseased.

3.11 Leukemia

Leukemia is one kind of cancer disease with white blood cells. It is state of the uncontrolled abnormal malignant proliferation of hemopoietic stem cells, which causes progressively increasing infiltration of the bone marrow and secondarily flood the circulating blood and other organs. Gum bleeding, thrombocytopenia, epistaxis, and lymphadenopathy are the symptoms of leukemia.

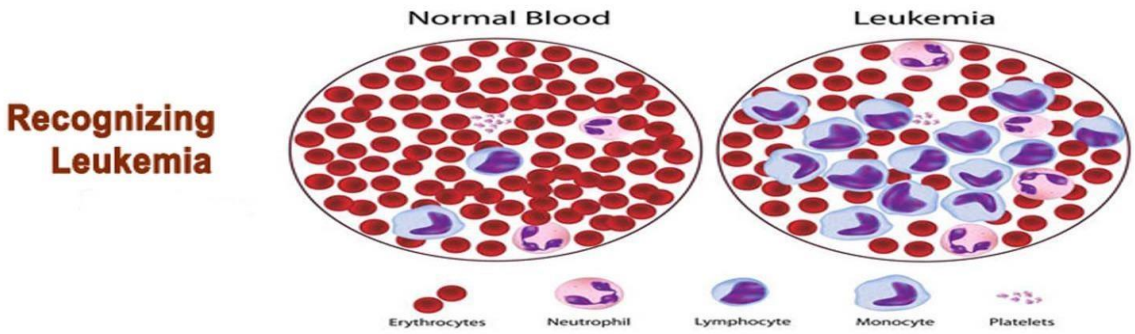

Figure 3. 12: Comparison of normal and leukemia diseased blood cell.

3.12 Malaria

Malaria is a viral life-threating illness. This disease is transmitted by certain mosquitoes. There are seven main symptoms can be shown on the patient like fever, profuse sweating, convulsion, hypoglycemia, shaking chills, jaundice, anemia.

Figure 3. 13: Malaria transmitted by mosquito.

3.13 Myocardial Infarction (MI)

Myocardial infarction is a state of imbalance oxygen supply and demand in the heart. After happening myocardial infarction patient may experience chest pain, shortness of breath, and dizziness.

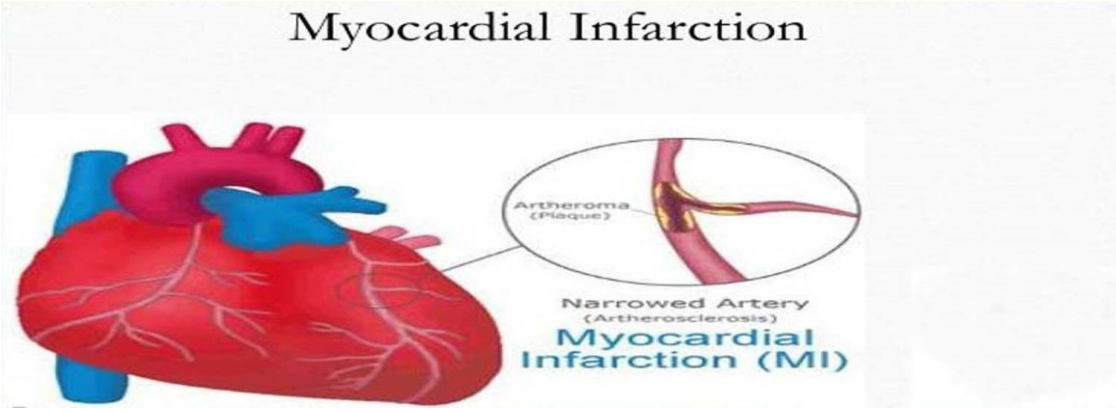

Figure 3. 14: Myocardial infarction.

3.14 Peptic Ulcer

Peptic ulcer is an acute or chronic disease. Usually, it can occur in any portion of the gastrointestinal tract that exposed to aggressive action of peptic juices. It is a very common disease and shows some main symptoms like heartburn, haemoptysis, epigastric pain, and abdominal pain.

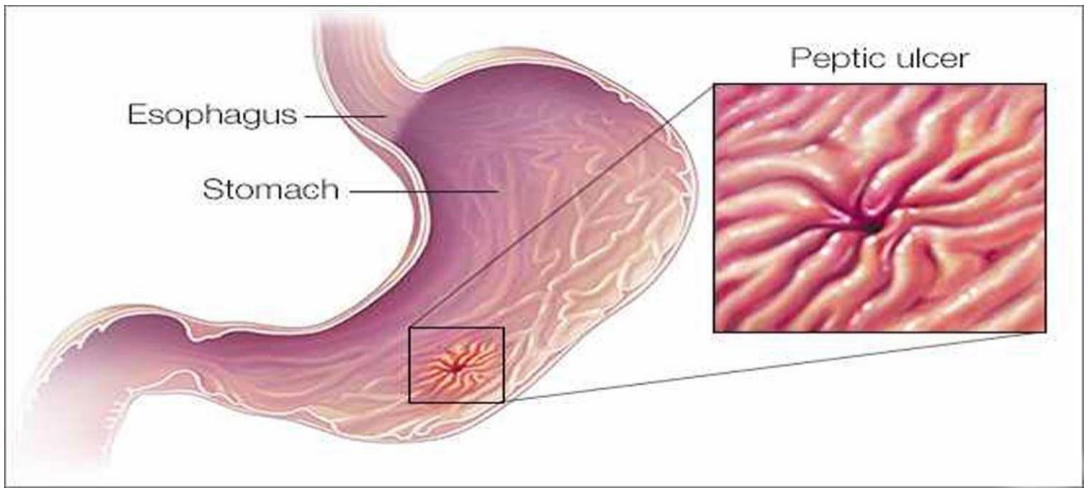

Figure 3. 15: Peptic ulcer.

3.15 Pneumonia

Pneumonia has recently been defined as a severe inflammation of the low respiratory (lung fires) associated with the improved radiation lung galaxy. This is a childhood common illness and symptoms are fever, cough, wheeze and shortness of breathing.

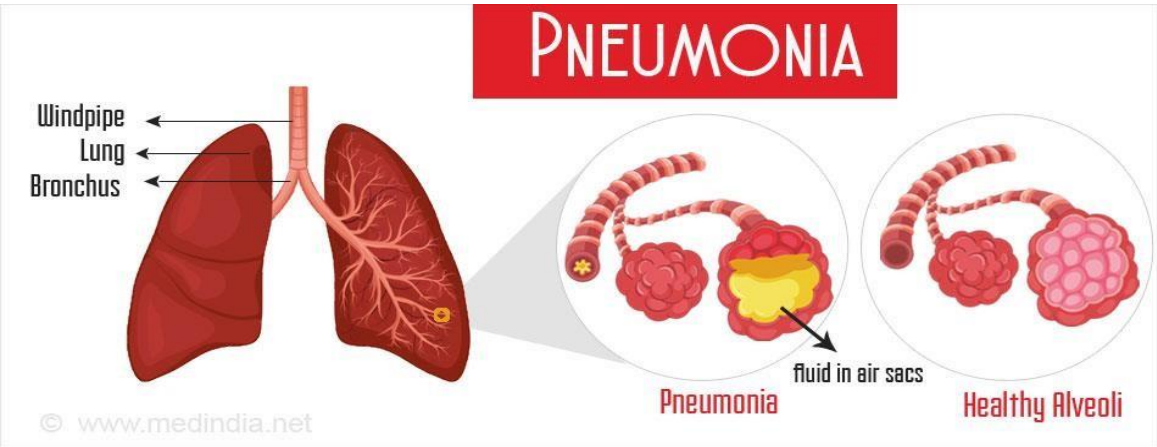

Figure 3. 16: Pneumonia

3.16 Rheumatic Fever

It's an inflammatory disease that can be developed as an unexpectedly treated staph neck or red hot fever problem. Streptio throat and red fever are caused by the transmission of stratococcus bacteria [41]. It has syndromes like Fever, Migratory Polyarthritis, and Red, hot, swollen joints.

Figure 3. 17: Rheumatic fever

3.17 Scurvy

Scurvy is the clinical manifestation of vitamin-C deficiency. Gum Bleeding, Angular Stomatitis, Hypothermia, Joint bleeding are the syndromes of scurvy.

Figure 3. 18: Scurvy

3.18 Stroke

Stroke is a very critical medical condition which paralyzes the organs of the body. Sometimes it can be very deadly if it cut down the blood circulation of the heart. Stroke is a medical emergency and emergency treatment essential. As soon as a person gets treatment for stroke, there may be less damage. Signs of mouth, arm weakness, and speech problems Symptoms of stroke [42].

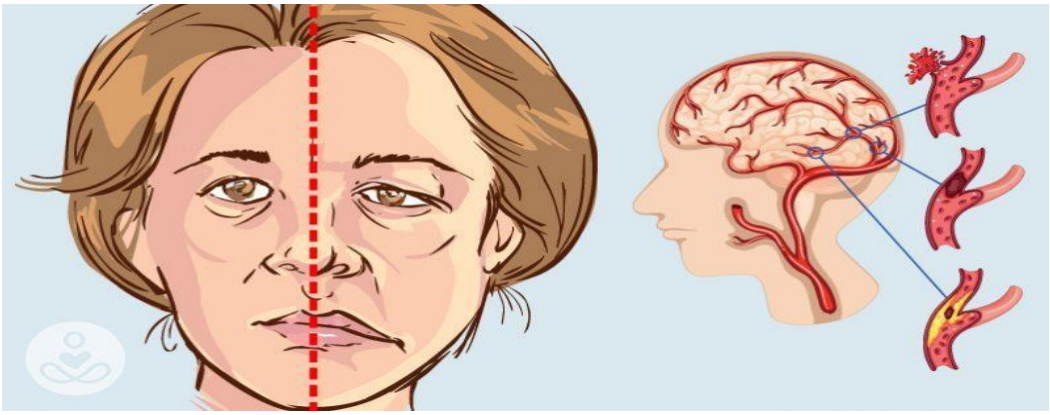

Figure 3. 19: Stroke

3.19 Tuberculosis

Tuberculosis is an infectious illness. It created for a bacteria named Mycobacterium tuberculosis. The main symptoms of tuberculosis are cough, haemoptysis, night sweats, weight loss, and fever and it is a spread by one person to another person breath air.

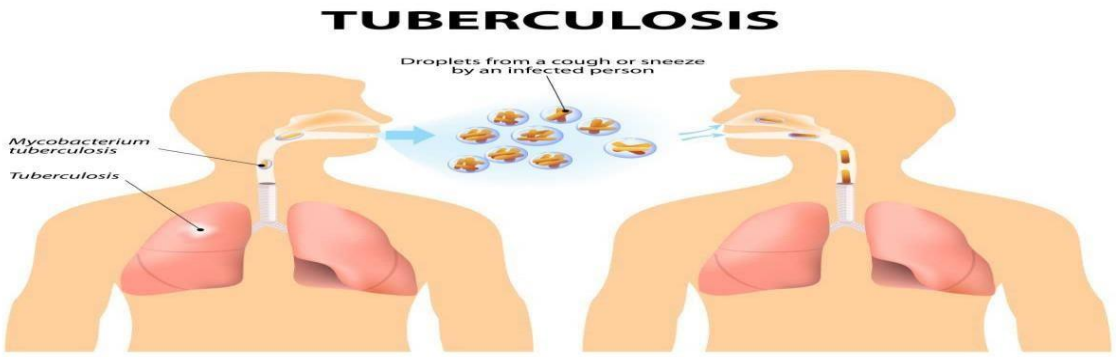

Figure 3. 20: Tuberculosis.

3.20 Typhoid Fever

Typhoid fever caused by bacterial infection that spread throughout the body, which affects many organs. Without proper treatment, it can be very fatal. It is caused by salmonella typhis, which is related to bacterial salicylate toxicity. [43] Signs can be seen as fever, stomach ache, diarrhea, constipation and lungs.

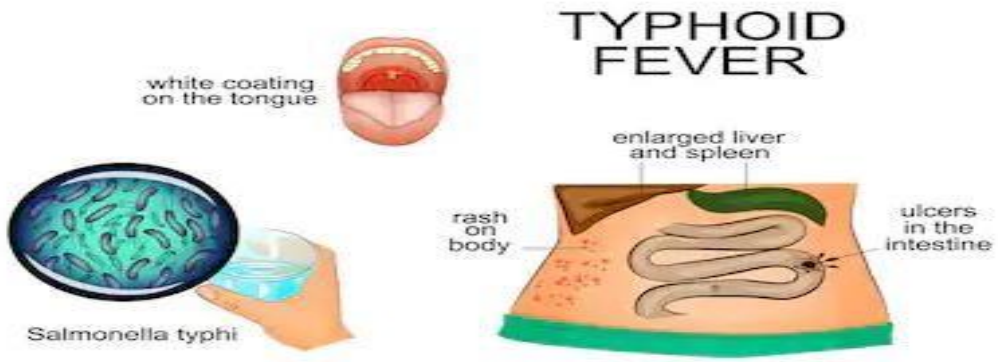

Figure 3. 21: Typhoid fever.

CHAPTER 4

Research Methodology

In this research work we are analyzing various kind of syndromes to identify their corresponding diseases. And we used data mining technique to find out patterns in syndrome dataset. Data will be processed and base on the outcome we tried to identify the corresponding diseases with desired efficiency.

4.1 Research object and instrumentation

The purpose of this research is to effectively analyze the possible result of the patient's dataset. To apply a prediction method, some features were developed in a model to identify the disease based on the characteristics of diseases. The steps in the process are as follows:

4.1.1 Filtering statistically significant features

To build an optimized machine learning model it is always vital to determine which independent attributes or features has more statistical significance. Approach that has taken to create the learning model is as follows:

a) Feature Selection:

To find out highly significant features there are five methods available and they are All in, Backward elimination, Forward selection, Bidirectional elimination and Score comparison respectively. All in is not a technical term but I used it demonstrate the idea where we intent to keep all of our attributes or features to build a machine learning model which is not efficient and contains a high probability to build a ‘garbage in - garbage out’ model. We have used ‘Backward elimination’ to find highly significant features because it is easy to implement and faster than all other methods available.

Steps involved in ‘Backward elimination’ are as follows:

Step 01: Select a significant level (SL) to qualify an attribute as highly significant.

Step 02: Fit the data with all possible attributes.

Step 03: If the value of P (Probability) is greater than SL, then go to Step 4 otherwise finish the process.

Step 04: Remove the attribute.

Step 05: Fit the data with rest of the attributes.

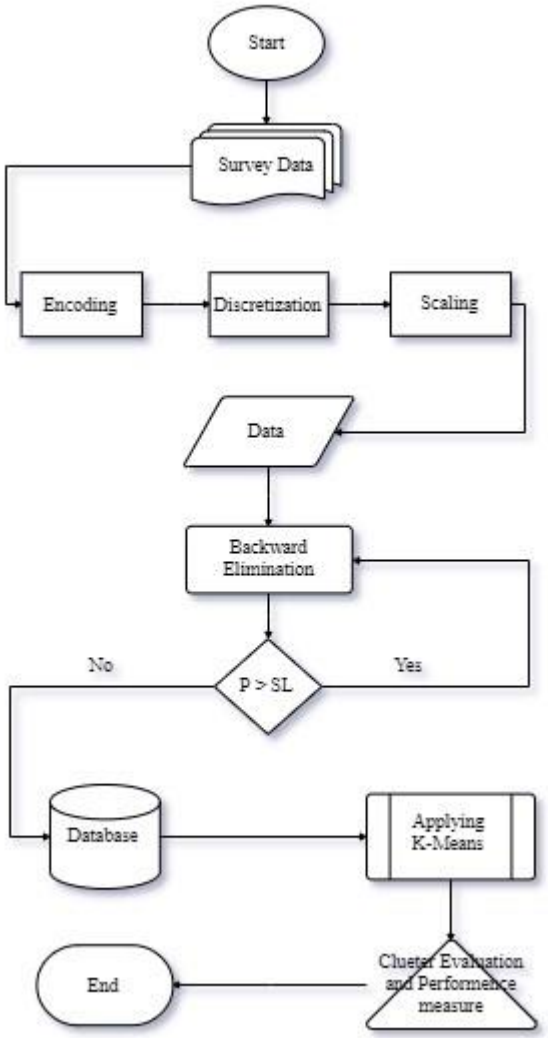

Figure 4. 1: Methodology

4.1.2 Applying K-Means and evaluation of clusters

When we have our dataset ready with highly significant features we have applied K-Means to the dataset to predict the outcome of each point based on every independent features that we have selected.

After K-Means finishes clustering, we have generated a clustering report to evaluate the performance of the model and we have taken ‘Accuracy score’ and ‘f1-score’ into consideration as numerical performance factor. The reason we took ‘f1-score’ over precision and recall is that the data we have used in this study is imbalanced and what ‘f1-score’ does is it combines precision and recall as harmonic means which punishes extreme values and gives a better intuition about the performance of the operation.

CHAPTER 5

Experimental Results and Discussion

In this study, we used an authenticated survey data for analysis and research purpose. The data were analyzed by the similarity measure (Calculated by "Euclidean distance") between data points and their cluster centroids. We have plotted data points before and after analyzing to show an abstract of operation. Data was optimized into two-dimension using t-SNE (t Stochastic Neighborhood Embedding) to plot in a 2D graph. Every cluster is represented by different colors and legends are used to identify them. A classification report is presented to give an insight on performance such as accuracy score, precession, recall, f1-score, etc.

5.1 Plotted Dataset

The scattered disease dataset is loaded by the python 'pandas' library and plotted by 'seaborn' for visualization.

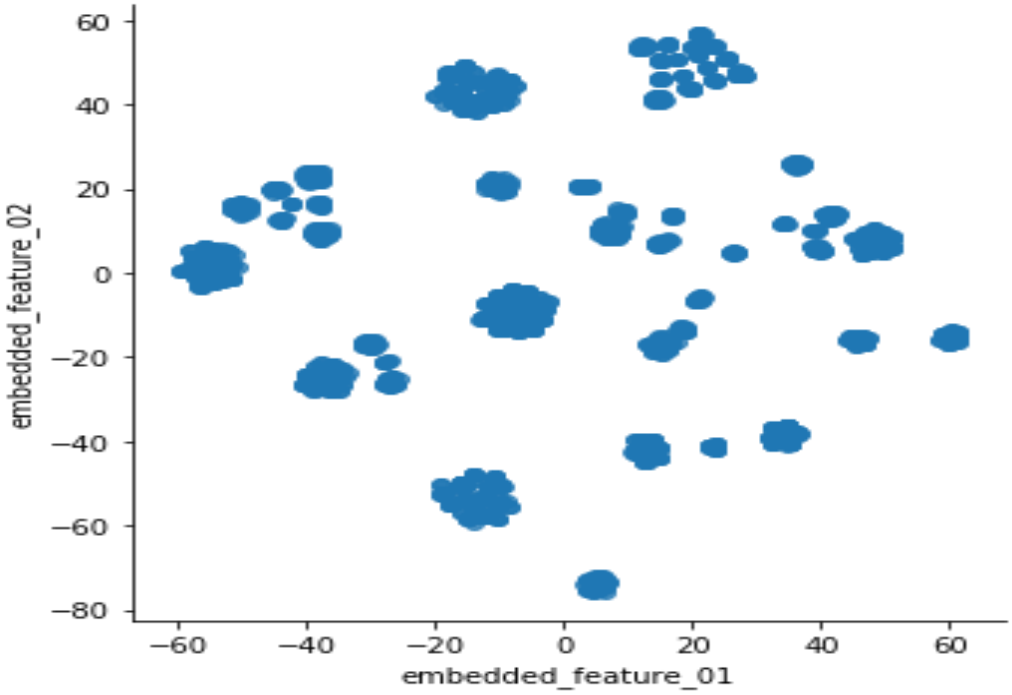

Figure 5. 1: Data points (Embedded in 2D)

5.1.1 Applying K-means and plotting clustered data

After applying K-Means clustering with k=18 the dataset is divided into 18 clusters. The regions are divided by measuring the distances between data points and centroids. Fig.5.1 shows the cluster with color codes that are formed after applying K-Means.

The original data points and predicted data points are shown using color codes below:

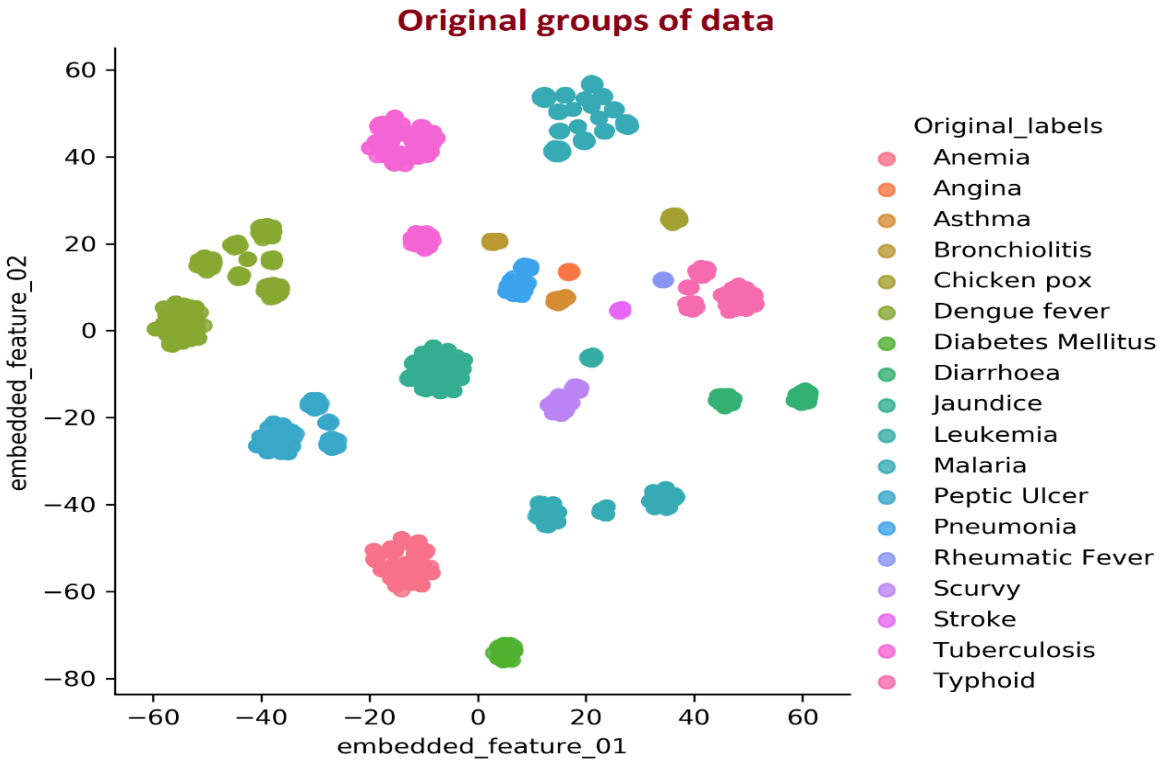

Figure 5. 2: Original Data points with color code

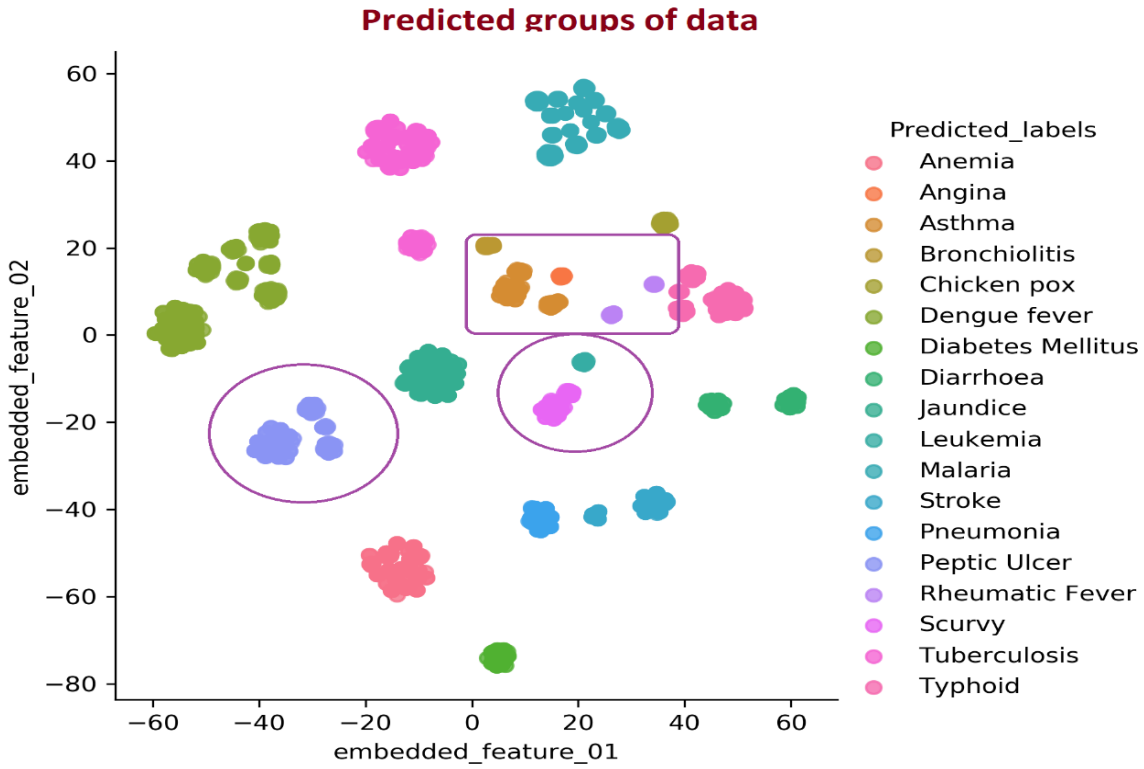

Figure 5. 3: Predicted clusters with color code

Figure 5.3 shows that some data points remain un-clustered or wrongly clustered. Some post optimization techniques can be applied to optimize the result further more. But we have considered it because of its performance and promising result. But it can be pointed as a future scope.

5.2 Clustering report

```
Result analytics:
==================================
Confusion matrix:
[[101   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0  11   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0  19   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0  14   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0  22   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0 202   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0  37   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0  61   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0 103   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0  14   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0 144   0  48   0   0  64   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0 111   0   0   0   0   0   0]
 [  0   0  50   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   8   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0  51   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0  11   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 143   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 101]]
----------------------------------
Clastering report:
                  precision    recall  f1-score   support

           Anemia       1.00      1.00      1.00       101
           Angina       1.00      1.00      1.00        11
           Asthma       0.28      1.00      0.43        19
    Bronchiolitis       1.00      1.00      1.00        14
      Chicken pox       1.00      1.00      1.00        22
     Dengue fever       1.00      1.00      1.00       202
Diabetes Mellitus       1.00      1.00      1.00        37
         Diarrhoea       1.00      1.00      1.00        61
          Jaundice       1.00      1.00      1.00       103
          Leukemia       1.00      1.00      1.00        14
           Malaria       1.00      0.56      0.72       256
     Peptic Ulcer        1.00      1.00      1.00       111
         Pneumonia       0.00      0.00      0.00        50
  Rheumatic Fever        0.42      1.00      0.59         8
           Scurvy        1.00      1.00      1.00        51
           Stroke        0.00      0.00      0.00        11
     Tuberculosis        1.00      1.00      1.00       143
           Typhoid       1.00      1.00      1.00       101

         accuracy                           0.87      1315
        macro avg       0.82      0.86      0.82      1315
     weighted avg       0.94      0.87      0.89      1315

Accuracy score: 86.84 %
==================================
```

Figure 5. 4: Confusion matrix and clustering report

5.3 Comparative performance measure with 'Agglomerative Hierarchical Clustering'

	Precision	Recall	F1-Score	Accuracy (In %)
K-Means Clustering	0.94	0.87	0.89	86.84
Agglomerative Hierarchical Clustering	0.94	0.79	0.82	78.71

Table 5.1: Comparison between K-Means and Agglomerative Hierarchical Clustering

It can be inferred from Table 5.1 that K-Means does a very good job in compare with hierarchical clustering containing almost 87% accuracy and 0.87 F1-score. And hierarchical clustering carries some bigger issues while working with big data. Some of the evident disadvantages are, hierarchical clustering is high in time complexity, the order of the data has a great impact on the outcome and very sensitive to outliers which is not good in this type of study.

CHAPTER 6

Conclusion, limitations and future research

6.1 Conclusion

To discover a new range of pattern and information using K-means clustering, we present this study to analyze the disease in the healthcare domain. We represent this study to analyze diseases in the healthcare domain to discover a new range of patterns and information using K-means clustering. This data mining-based system reduces the impact of human error and enhances the effectiveness of the analysis. The K-Means algorithm can be quite useful in data mining or predictive analysis of diseases and can produce a promising result but does not produce the most accurate results as desired to make effective decisions. To make K-Means more effective it can be used in combination with other algorithms to produce accurate, relevant and useful results. That being said, the learning model we have proposed in this study will be handy enough for physicians and medical practitioners to effectively predict hazardous cases and evaluate accordingly.

6.2 Limitations

- In primary test cases we skip some diagnosis factors such as ‘gender’, ‘age’, ‘Previous medical records’ so on.
- This approach mostly rely on working dataset. More accurate and larger amount of data, more accuracy in result.

6.3 Future scope

- The proposed model can be compared with other data mining techniques specially supervised learning techniques to check reliability.
- Another classifiers like SVM, naïve bias etc. can be used to reduce the complexity and increase the accuracy.

CHAPTER 7

References

[1]Usama Fayyad, Gregory Piatetsky Shapiro and padhraic Symyh, “The KDD Process for Extracting Useful Knowledge from Volumes of Data”, Communication of the ACM, Vol. 39, No. 11, pp. 27-34,1996.

[2]Chauhan R, Kaur H, Alam M A, “Data Clustering Method for Discovering Clusters in Spatial Cancer Databases”, International Journal of Computer Applications , (0975 – 8887) Vol.10– No.6, November 2010.

[3]AmandeepKaurMann ,NavneetKaur ,”Survey Paper on Clustering Techniques “Volume 2, Issue 4, April 2013 ISSN: 2278 – 7798.

[4]Jain A.K., Murty M.N., and Flynn P.J., “Data Clustering: A Review”, ACM Computing Surveys, 31 (3). pp. 264-323, 1999.

[5] C. C. Aggarwal, J. Han, J. Wang, and P. S. Yu, "Aframework for projected clustering of high dimensional datastreams," in Proceedings of the Thirtieth internationalconference on Very large databases-Volume 30, 2004, p.863.

[6] R. Agrawal, J. E. Gehrke, D. Gunopulos, and P. Raghavan,"Automatic subspace clustering of high dimensional data fordata mining applications," Google Patents, 1999.

[7]SANDRO VEGA-PONS and JOSÉ RUIZ-SHULCLOPER, A SURVEY OF CLUSTERING ENSEMBLE ALGORITHMS, International Journal of Pattern Recognition and Artificial Intelligence, 2011, 337-372.

[8]Zenon Brzoza, Canonica Walter, Martin K Church and Martin Metz, The EAACI/GA2LEN/EDF/WAO Guideline for the definition, classification, diagnosis, and management of urticaria: the 2013 revision and update, John Wiley & Sons A/S. Published by John Wiley & Sons Ltd, 2014, 1-20

[9]RatnadipAdhikari and R.K. Agrawal, A Novel Weighted Ensemble Technique for Time Series Forecasting, Springer, 2012, 38-49.

[10]Sarwesh Site and Dr. Sadhna K. Mishra, A Review of Ensemble Technique for Improving Majority Voting for Classifier, International Journal of Advanced Research in Computer Science and Software Engineering, 2013, 177-180.

[11]M. ANBARASI, E. ANUPRIYA and N.CH.S.N. IYENGAR, “Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm, International Journal of Engineering Science and Technology”, 2010, 5370-5376.

[12]Nidhi Bhatla and Kiran Jyoti, “An Analysis of Heart Disease Prediction using Different Data Mining Techniques, International Journal of Engineering Research & Technology”, 2012, 1-4.

[13]Asha Rajkumar and Mrs. G.Sophia Reena, “Diagnosis Of Heart Disease Using Datamining Algorithm, Global Journal of Computer Science and Technology”, 2010, 38-43.

[14]Mai Shouman, Tim Turner and Rob Stocker, “Using data mining techniques in heart disease diagnosis and treatment”, IEEE, 2012, 189-193.

[15]Jyoti Soni, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction”, International Journal of Computer Applications, vol.17, pp. 43–48, Mar. 2011.

[16]B.M.Patil, Ramesh C.Joshi and Durga Toshniwal, “Effective framework for Prediction of Diseases outcome using medical Datasets clustering and classification”, Published in International Journal of computational Intelligence studies, Vol 1, Issue 3, pages 273-290, August 2010.

[17]Sellappan Palaniappan, Rafiah Awang “Intelligent Heart Disease Prediction System Using Data Mining Techniques”Department of Information Technology Malaysia University of Science and Technology Block C, Kelana Square, Jalan SS7/26 Kelana Jaya, 47301 Petaling Jaya, Selangor, Malaysia.

[18]"CSV File Reading and Writing" (http://docs.python.org/library/csv.html).. Retrieved July 24, 2011. "is no "CSV standard"".

[19]Y. Shafranovich. "Common Format and MIME Type for CommaSeparated Values (CSV) Files" (http://tools.ietf.org/html/rfc4180) Retrieved September 12, 2011.

[20]home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html “A tutorial on clustering algorithms”.

[21]Shadab Adam Pattekari and Asma Parveen “Prediction System For Heart Disease Using Naïve Bayes” International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. Vol 3, Issue 3, 2012, pp 290-294.

[22]S. Vijayarani and S. Sudha, “An Efficient Clustering Algorithm for Predicting Diseases from Hemogram Blood Test Samples”, Indian Journal of Science and Technology, vol.8, pp. 1–8, Aug. 2015.

[23]Jyoti Soni, “Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction”, International Journal of Computer Applications, vol.17, pp. 43–48, Mar. 2011.

[24]K.Rajalakshmi, Dr.S.S.Dhenakaran and N.Roobini “Comparative Analysis of K-Means Algorithm in Disease Prediction”, International Journal of Science, Engineering and Technology Research, vol.4, pp. 2697–2699, Jul. 2015.

[25]Shantakumar B.Patil Y.S.Kumaraswamy “Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network” European Journal of Scientific Research ISSN 1450216X Vol.31 No.4 (2009), pp.642-656.

[26]Bala Sundar V,T Devi, N Savan,”Development of a Data Clustering Algorithm for Predicting Heart”, International Journal of Computer Applications (0975 – 888) Vol.48– No.7,2012.

[27]Sachin Shinde, Bharat Tidke, “Improved K-means Algorithm for searching Research Papers”, International Journal of Computer Science & Communication networks, ISSN: 2249-5789, Vol.4 (6), 197202.

[28]M.Umamaheswari, Dr. P. Isakki @Devi, “Myocardial Infarction Prediction using K-means Clustering Algorithm”, International Journal of Innovative Research in Computer and Communication, Vol. 5, Special Issue 1, March 2017.

[29]Muhammad Zulfadhilah, Imam Riadi, Yudi Prayudi, “Log Classification using K-means Clustering for Identify Internet User Behaviours”, International Journal of Compiler Applications, (0975-8887), Vol.154No.3,November 2016.

[30]Oyelade, O. J, Oladipupo, O. O and Obagbuwa, I. C, "Application of k-Means Clustering algorithm for prediction of Students' Academic Performance", International Journal of Computer Science and Information Security, Vol. 7, o. 1, 2010.

[31]K.Rajalakshmi,Dr.S.S.Dhenakaran,N.Roobin"Comparative Analysis of K-Means Algorithm in Disease Prediction", International Journal of Science, Engineering and Technology Research (IJSETR), Vol. 4, Issue 7, July 2015.

[32]A. K. Pandey, P. Pandey, K. L. Jaiswal, and A. K. Sen, "Data Mining Clustering Techniques in the Prediction of Heart Disease using Attribute Selection Method," International Journal of Science, Engineering and Technology Research (IJSETR), ISSN: 2277798, Vol 2, Issue10, October 2013.

[33]R. Das, I. Turkoglu, and A. Sengur, "Diagnosis of valvular heart disease through neural networks ensembles," Elsevier, 2009.

[34]M. Karaolis, J. A. Moutiris, and C. S. Pattichis, "Association rule analysis for the assessment of the risk of coronary heart events," Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2009.

[35]S. Shilna and E. Navya, " Heart disease forecasting system using kmeans clustering algorithm with PSO and other data mining method," International Journal On Engineering Technology and Sciences (IJETS™), ISSN(P): 2349-3968, ISSN (O): 2349-3976, Vol. 3, Issue 4, April 2016.

[36]K.R. Lakshmi, M. V. Krishna and S. P. Kumar, "Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability," International Journal of Scientific and Research Publications, ISSN 2250-3153, Vol.3, Issue.6, June 2013.

[37]K. Solanki, P. Berwal and S. Dalal, "Analysis of application of data mining techniques in healthcare," International Journal of Computer Applications, Vol. 148, No.2, August 2016.

[38]M. Verma, M. Srivastava, N. Chack, A. K. Diswar and Nidhi Gupta," A Comparative Study of Various Clustering Algorithms in Data Mining, "International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue 3, pp. 1379-1384, 2012.

[39]S. Revathi and T. NalinI, "Performance Comparison of Various Clustering Algorithm," Vol. 3, Issue 2, ISSN: 2277 128X, February 2013.

[40]R. Chauhan, H. Kaur and A. Alam, "Data Clustering Method for Discovering Clusters in Spatial Cancer Databases," International Journal of Computer Applications, Vol. 10, No.6, November 2010.

[41]https://www.mayoclinic.org/diseases-conditions/rheumatic-fever/symptoms-causes/syc-20354588

[42]https://www.nhs.uk/conditions/stroke/

[43]https://www.nhs.uk/conditions/typhoid-fever/