Region Based Disease Prediction Using Data mining and Machine learning.

BY:

Tuhedul Islam

ID: 152-15-545

And

Md. Nahid-Al-Mamun

ID: 152-15-531

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering (B.Sc. In CSE).

Supervised BY:

Sheikh Abujar

Lecturer, Department Of CSE

Daffodil International University

Co-Supervised By:

Dewan Mamun Raja

Lecturer, Department Of CSE

Daffodil International University

Daffodil International University

Dhaka, Bangladesh

September 2018

APPROVAL

This Thesis title "**REGION BASED DISEASE PREDICTION USING DATA MINING AND MACHINE LEARNING**" submitted by Md. Nahid Al Mamun ID No: 152-15-531 and Tuhedul Islam ID No: 152-15-545 to the department of computer science and engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science and Engineering (BSc) and approved as to its style and contents. The presentation has been held on 6 APRIL 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**

Associate Professor & Associate Head

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

_____

Saif Mahmud Parvez **Internal Examiner**

Lecturer

Department of Computer Science and Engineering

Faculty of Science & Information Technology

Daffodil International University

_____

Dr. Mohammad Shorif Uddin **External Examiner**

Professor

Department of Computer Science and Engineering

Jahangirnagar University

DECLARATION

We hereby declare that, this research has been done by us under the supervision of **Sheikh Abujar, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this research nor any part of this research has been submitted elsewhere for award of any degree or diploma.

Supervised by:

_____

(Sheikh Abujar)
Lecturer
Department of CSE
Daffodil International University

Co-Supervised by:

_____

(Mr. Dewan Mamun Raza)
Lecturer
Department of CSE
Daffodil International University

Submitted by:

_____

(Md. Nahid-Al-Mamun)
ID: -152-15-531
Department of CSE
Daffodil International University

_____

(Tuhedul Islam)
ID: -152-15-545
Department of CSE
Daffodil International University

ACKNOWLEDGEMENTS

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year research successfully.

We really grateful and wish our profound our indebtedness to **Sheikh Abujar**, **Lecturer**,

Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & extreme interest of our supervisor in the field of "*REGION BASED DISEASE PREDICTION USING DATA MINING AND MACHINE LEARNING*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice ,reading many inferior draft and correcting them at all stage have made it possible to complete this project.

We would like to express our heartiest gratitude to **Mr. Dewan Mamun Raza**, **Lecturer**, Department of CSE Daffodil International University and **Dr. S.M Aminul Haque, Assistant Professor & Associate Head**, Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University. We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work. Finally, we must acknowledge with due respect the constant support and patients of our parents.

April 6, 2019

ABSTRACT

Day by day increasing the value of time of human, that's why human want to try analyses all the data related with life and environment, for making future life easy and comfortable. In this time, worldwide uses data mining technique and machine learning technique. Data mining technique is used for discovering information and knowledge from large database and find the pattern and some possibly unexpected pattern in between different data. Machine learning technique give ability to the program to learn automatically. The medical industry uses data mining and machine learning first priority for disease prediction and gain some unexpected knowledge by compering huge data. The medical industry collect huge amount of healthcare data. Sad to say that they do not discover hidden relationship for this data .so here advanced data mining technique is very helpful. Using data mining and machine learning technique to doing this research is a prototype for disease prediction, by conducting the appropriate use of biological profile.

In this paper, our prediction will describe about a disease when or where it will occur or epidemic from region to region. If we make a better prediction then people have to conscious about health in a region.

LIST OF FIGURES

LIST OF ABBREVIATIONS

WORD	ABBREVIATIONS
ML	Machine Learning
MLT	Machine Learning Technique
DM	Data Mining
LRA	Linear Regression Algorithm
K-NN	K- Nearest Neighbor
CNN	Convolutional Neural Network
WHO	World Health Organization
CVD	Cardiovascular Disease
WLE	World Life Expectancy
PR	Polynomial Regression

TABLE OF CONTENTS

CHAPTER 3: RESEARCH METHODOLOGY

CHAPTER 4: EXPERIMENTAL RESULT AND DISCUSSION

CHAPTER 5: CONCLUSION

CHAPTER 1
INTRODUCTION

1.1 Introduction

In medical industry data mining and machine learning is mainly use for diseases pattern analysis. In this paper data mining and machine learning is used for diseases prediction. Different types of diseases outbreak are often happens [1]. In 2002 44% people death by chronic diseases, 23% people death by cardiovascular diseases, 6977 people affected by dengue with 17 death in 2018 etc. If human predict diseases outbreak time, then they are make some action against this diseases as soon as possible. In this paper by analyzing data to produce result about diseases outbreak time [2]. Machine Learning using the past experience like as human brain a system can learn itself by analyzing huge data and can predict about a new thing. Here use data mining and machine learning technique to predict that disease which will come first in a country and spread up [1]. Using the way, we will predict the death rate, harmed rate affected by the disease and also provide the information about the next disease outbreak (already discovered or not discovered by medical science) appear before in that country.

1.2 Objective

This paper predominantly pursuit on country benefit. When a country is very alert about future diseases, it's very helpful. Cause they can process there current healthcare data and gain some result about future disease outbreak time, finally planning some action against future diseases. Now almost every country people are alert, even people are thinking about great future for next generation. So, this paper playing significant rule globally to safe human.

1.3 Rationale of the study

In this paper, we will try to predict about a disease which comes first in the different region and make it expansion in that region. So the rationale of the study is predictive which predict about a disease in different region. To do the work, different machine learning algorithm and data mining technique we have use [3]. The prediction of different diseases according to our research question will make the people conscious in any region and it will be much helpful about their healthcare. Machine learning algorithms are used for that prediction.

1.4 Research Question

Any research paper must have a research question and that explain about the thing whatever they want to do in their research. So we have also a question what we want to do in this paper. Based on the question we will predict about a disease which will first appear in a region.

The Research question is “When a disease first appear in a region and create epidemic or spread up in that region”.

According to the question, let a disease first appear in a region and when it will make outbreaks in other region.

1.5 Expected Output

Prediction is not always accurate and sometimes it will give inaccurate result in different situation by analyzing the data. The accuracy of a prediction may be good, best or worst.

From this work, our expectation is to conscious about the health of people in different region in worldwide. If the people in a region can know about a disease when it will outbreaks on that region then they have to take necessary steps to get read of that disease and be happy in their normal life or they lead a peaceful and blissful life in the future world.

CHAPTER 2
BACKGROUND

2.1 Introduction

The title we have selected ‘’Region Based Diseases Prediction Using Data mining And Machine Learning ‘’ is basically means when does any disease first appear in a region to region.

The aim of this paper is to predict diseases which is region based. Likely, a diseases have already discovered in a country and when it will spread up in another country in the world. How much time a diseases takes to appear in one country to another country [4]. By identifying the different diseases appear in a region, the system will predict the next disease which will spread up on that region.

It is first time, we are trying to suggest a system which will predict a diseases when it will spread up on region based using data mining and machine learning.

2.2 Related Works

There are so much research paper in the different journals about disease what already discovered in the medical science. All time the specialists of the medical science in the world are trying to solve the problem about any single disease and the researcher are trying to make better the prediction about it.

Our work in this paper, we make a prediction about a disease when it comes appear before in a region and occurs outbreak of the disease. But there is not enough research like the research we are working for and make a prediction.

In the different online journals, the published papers are shown in the statics of death rate, partially death rate between male and female many more things but the research we are working on is different based on the prediction.

2.3 Research Summary

Cancer, Hepatitis, and cardiovascular diseases are the most serious and diverse diseases in the world and occur epidemic in different country. There are so much research about these kind of disease and these paper find out the different statics like percentage, severity for male and female of the total population in a region [5].

World Health organization (WHO) statics say that 17.5 million people died from Cardiovascular disease (CVD) in 2012 which representing 31% of all global deaths. One person kills by heart disease every 34 second in USA [6]. Unfortunately, all Doctors do not possess experts in every sub specialty and lack of resource person for particular area. Hepatitis disease is from the most dangerous disease that is an inflammation of the liver and cause of death in the world specially Hepatitis C. It may occur by hepatitis virus or by using many types of drug and alcohol and it is a contagious liver disease [7]. Every year, three or four (3-4) million people are infected with the hepatitis C virus and 350,000 people die from hepatitis related liver disease. The five most common sites of cancer which is diagnosed in 2012 were lung, prostate, colorectal, stomach, and liver cancer [8]. Among women the five most common cancer diagnosed and they were breast, colorectal, lung, cervix, and stomach cancer. Cancer that means the abnormal growth of cell that may occurs any part of human body and break down the normal activity of cell. Day by day the new cases of cancer are increasing in various region and the newly discovered cases are so much detrimental.

Neoteric cases of different virus and bacteria are unfolded in the modern world depending on the environment of a region to region. Data mining and machine learning used in medical science area is highly potential to exploring hidden relationship, knowledge or pattern in the data set. Huge number of data are available of medical profile of different region's people [9].

According to another research study, heart disease is the threat of current world and also for the future world. It increases day by day and different region's people are attacked in their everyday life due to their livelihood [10]. Not only in the USA but also all over the country in the world are affected by different types of heart disease and new cases of that disease are increasing highly over the world.

Neural disorder of human brain are widely causes in different region. Parkinson disease is one of the disease of neural disorder which poses the abnormal behavior of human.

Another dangerous disease is the Tuberculosis (TB) which attack on the human lung and also damage the other part of body of human. In different country it causes epidemic and occurs the harms in region to region [11].

Not only the communicable disease but also the non-communicable disease and its different factors are spread up in worldwide from region to region. Our attempt is to ascertain when or where any disease will epidemic or first occur.

2.4 Scope of problem

When a disease becomes epidemic or outbreaks in a region then the people start to try for prevention and takes the proper steps to get read of the disease. But the people get informed by any system that will describe about a disease when it will arrive in a region then they can follow the efficient way to solve it and rises the consciousness about that dangerous problem before it happens.

Most of the hospital's doctors make decision depending on their previous experience whatever they gather. For that reason, there are more lacks on that decisions but if the doctors use the hidden knowledge of medical data within a system as their companion then the decision may be declared as most as proper.

2.5 Challenges

The most important thing is the proper medical data of different region depending on which our prediction model will make a decision. If the data set is not real life related then the decision may be poor. Depending on the algorithm accuracy, the decision of prediction model is vary. So choosing the algorithm is one of the important challenge.

CHAPTER 3
RESEARCH METHODOLOGY

3.1 Introduction

For prediction based research, many machine learning (ML) algorithm are used to predict the expected outcome. All the regression algorithms can be used to make prediction model and also the classification algorithms can be used. The regression algorithm are vary from depending on the attributes. Like the Linear regression algorithm, Polynomial regression algorithm, multiple linear regression algorithm and many more. In Regression algorithm, the most important thing is about categorical value which is not supported. The classification algorithm like K-NN may be used for prediction. In Naive Bayes, the categorical values are supported.

3.2 Data Collection Procedure

Actual and real life data make the prediction much efficient and a powerful predictive model. Data collection is the most important thing for machine learning and data mining to gather a knowledge or pattern from huge amount of data that will help the system to learn new thing. Using the discovered pattern, a system can make a proper prediction and the prediction always helpful for medical science as companion of doctors.

Collection of data must be the related works whatever anyone want to do in a research paper. The source of our data is the online webpage like world health organization (WHO), world life expectancy (WLE), Wikipedia and many more.

Firstly our attempt was to collect the epidemic or spread up date of a disease in the different region. The collection of region based spread up disease discovery date is much tough because there is no report or proper data set or that is not recorded properly in any data repository.

The collected data set is prepared in the excel (.csv) file which is describe in a table below.

Table 3.1: Data set format

Country	Disease	Year
USA	Coronary Heart Disease	1913
Bangladesh	AIDS	1981
Australia	Cancer	1950

3.3 Statistical Analysis

Linear regression algorithm uses the linear equation for calculation. The algorithm divide the data set into two sets which one is training and another is testing data set. Firstly, the algorithm show the predictive linear regression line according to the training data set. And then use the testing data set for testing the model efficiency.

The equation,

$$y = mx + c$$

Here,

y = Dependent variable.

x = Independent variable.

c = Constant variable.

m = Slope.

After completing the data preprocessing part, using the linear equation algorithm gives the predictive value. In that case, data must be converted from categorical to numerical value.

3.4 Implementation Requirement

Many machine learning algorithm are already well designed for many more prediction which are available. Recently, artificial Intelligence has turn out prevalent and uses in different real life needs. In different sector's people are trying to make their life easier and flexible by applying AI. The scientists or the specialists are trying to use AI system as their work companion. For example, any business prediction to make their business more profitable, predicting future environment of a region or a world by analyzing the previous statistics. Now a days, economists are using AI to predict future market prices to make a profit, doctors use AI to classify whether a tumor is malignant, meteorologists use AI to predict the weather, HR recruiters use AI to check the resume of applicants to verify if the applicant meets the minimum criteria for the job. The inspiration behind such global use of AI is machine learning algorithms. Machine learning is under the control of supervised learning or unsupervised learning.

Linear Regression algorithm, Polynomial regression, Logistic regression, Classification and regression tree, k- nearest neighbors, Support vector machine, Naive Bayes all are the machine learning algorithms. Linear regression is an algorithm based on supervised learning. Linear regression defines a linear relationship between the input variable and the output variable. The algorithm is simple, which makes it a great place to start thinking about algorithms more generally. For the prediction result what we want in this paper, we have use linear regression algorithm. Analyzing the dependent variable the algorithm predict a regression line. By using the polynomial regression we also get another output.

About Regression techniques which is differ based on the number of independent variables and the other thing is that the type of relationship between the independent and dependent variables.

Simple linear regression is a type of regression analysis where the number of independent variables is one and there is a linear relationship between the independent(x) and dependent variable(y).

3.5 Flowchart and Model

Data mining is a process, it's a part of Artificial Intelligence. This technique work for identifying relationship, hidden pattern and knowledge by processing huge amount of data. Show some possibly unexpected result.

Machine learning is also a part of Artificial Intelligence. In this process machine can learn like human. In this case machine need training dataset, after analyze this dataset machine can produce result for test dataset.

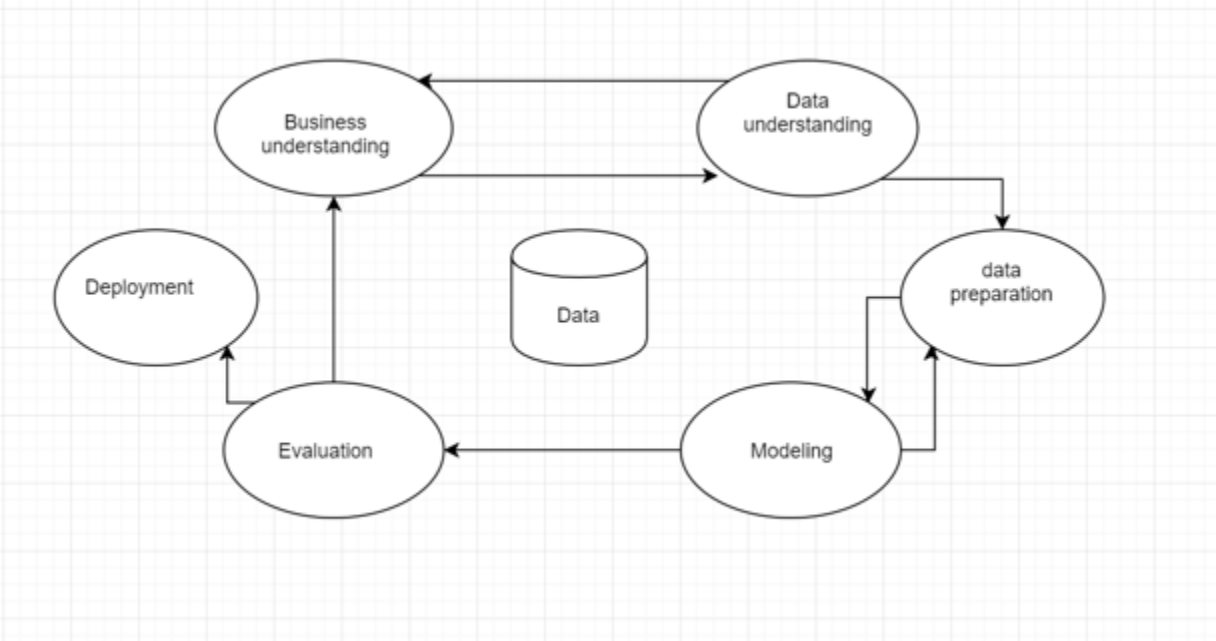

Figure 1: Data mining process model

Data mining process:

Here is description process about data mining process with the focus on the CRISP-DM model. While there are many other data mining data mining process including SAS Institute's Semih model. Which stand for sample explore modify model and assess. Here I will focus on the crisp methodology. It is also important to note that, this intro is simply an introduction the different

stage of data mining process model are actually way more detailed in practice. This model is popular methodology that provide a structured approach to planning a data mining project. The six phases in this data mining process.

Business understanding:

The first step in this stage is to understand the problem to be solved it really is an iterative process of discovery with the data understanding phase as you can see here. What you really want to accomplish in this stage is to figure out from business perspective and creativity often a massive role in this business understanding stage. This specific steps are number one determine business objectives, number two assess the situation which is really detailed fact finding exercise, number three the data mining goals ordain and number four the project plan consequence.

Data understanding:

In this stage data is required and before does anything with that data he or she first needs to understanding the strength and weak limitation of the data. Because rarely will the data exactly match the problem that he or she trying to solve. It also important to recognize that data cost money and some data may not be available so data scientist needs to evaluate the cost and benefits of all the different potential data sources. In this stage the specific step are number one collect initial data, number two describe the data number three explore the data and number four verify the quality of the data

Data preparation:

Essentially analytics technologies often require that data be in a form that's different from how the data is initially provided and conversion necessary. Some common data preparation examples include number one converting data to tabular format, number two removing missing value, number three converting data to different types and number four scaling numerical values. This stage is one of the most time consuming stage. In this stage the specific steps are number one select data, number two clean the data, number three construct the data and number four integrate the data.

Modelling:

Applied to the data with the output hopefully being some sort of model or pattern but it's really important to note that not all patterns are valid. Data comes to light new models are built and as model are built you may go back to the preparation stage, If you realize that the model just are not very telling. In this stage the specific steps are number one select the modeling technique, number two generate tests for model robustness, number three build the model and number four assess the model.

Evaluation:

This stage assess the data mining results both from quantitative and qualitative perspectives and have a keen attention to detail determine. If the result are both justifiable and feasible. All patterns are not valid because patterns really can be quote unquote found in any data. Its encouraged to do this before the deployment stage because it usually end up cheaper and safer than simply skipping

the stage and going directly from modeling to deployment. That you may go back to the business understanding stage, if you realize in the evaluation stage that what you have is not meaningful. In this stage the specific steps are number one evaluate the result, number two review the process and number three determine the next step.

Deployment:

It is last step in data mining process. This is where the result of data mining are put into really use. Data science teams pass a working prototype to the deployment teams. So that they can recorded for the production environment. In this stage the specific steps are number one plan the deployment, number two planning for the patronage and monitoring, three review the system and four construct the final report of values.

Project Model:

By analyzing the data, system will provide a predictive result. After analyzing data, if the system will get new pattern then system will save the data as learning point and make notification for medical science. Thus, the system will learn new thing day by day. The System is attached below.

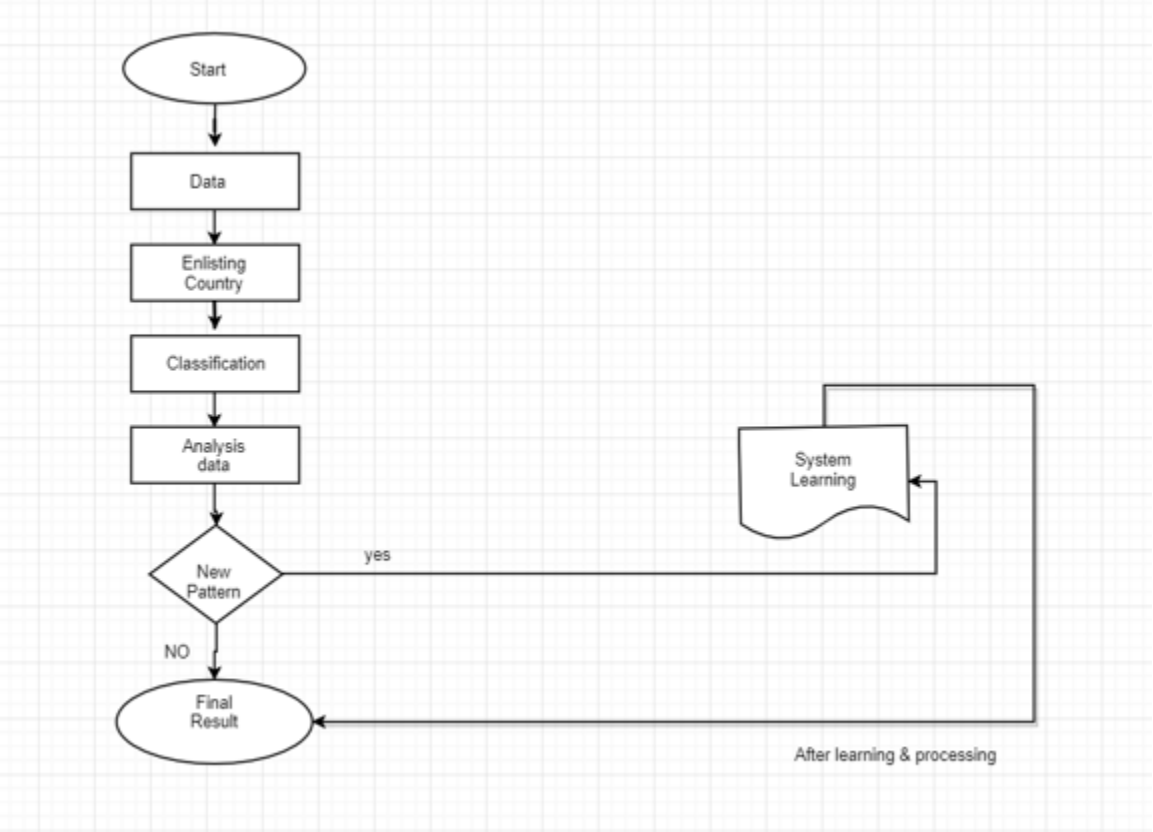

Figure 2: Flow Chart

CHAPTER 4
EXPERIMENTAL RESULT AND DISCUSSION

4.1 Introduction

Using the machine learning (ML) Linear Regression (LR) algorithm, we are trying to predict about a disease when or where it will appear as a new disease in a region. Polynomial regression algorithm also used for another output but the LR is mostly better.

The output of the prediction of training data set or the testing data set is shown below. Most of the collected data in the data set is used as training data set and the other data set is used as testing data set. If we divide the data set into training and testing then 70% of the data set are used as training and 30% data are used as testing.

4.2 Experimental Result

After applying the LRA on the data set then the graphical output provided by the algorithm is given below. For the training data set one output is provided and for the testing data set another output is provided.

Output of the training data set (LR),

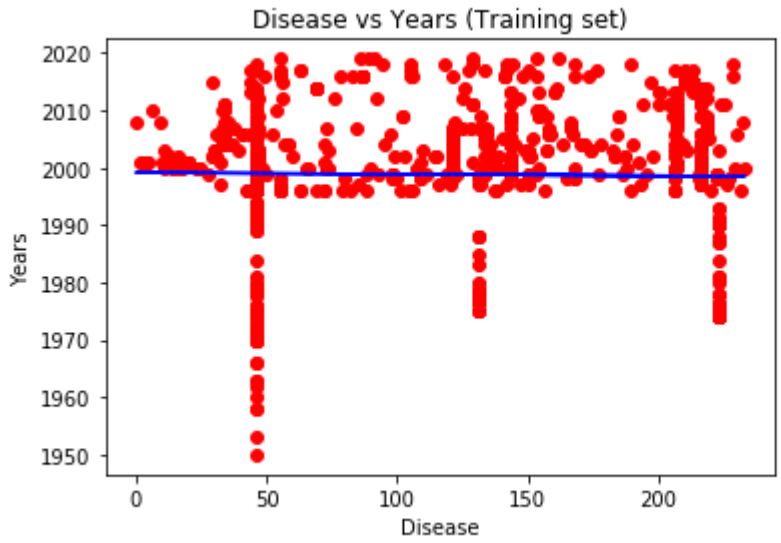

Figure 3: Training data set regression line (LR)

In the figure, the output is representing disease versus year. Here the trained regression line is blue color and predicted countries are the splitting red circle. System learn from the data set to predict the testing data set.

From the graph,

Let the numerical value against a disease is 49 (forty nine) and the year 1950 is set up against a country which is one of the splitting red circles shown in graph. It actually means when a disease arrive in a region according to trained data set.

Output of the testing data set (LR),

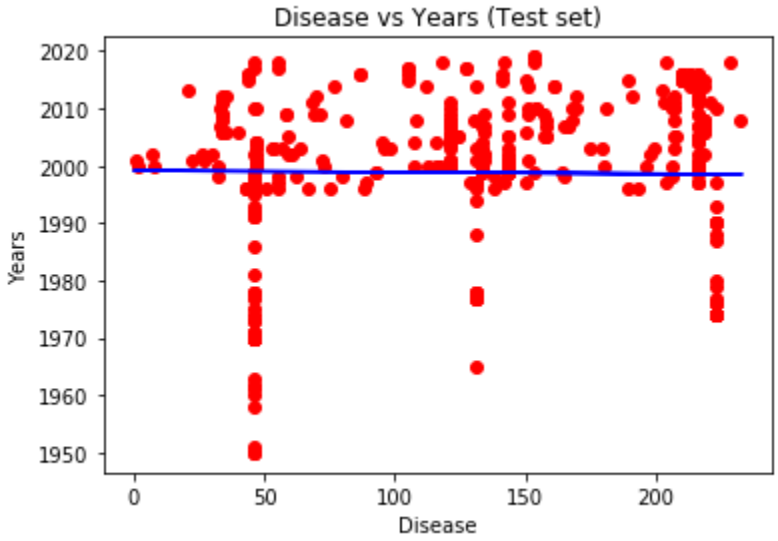

Figure 4: Testing data set regression line (LR)

When the model is trained after completing all the process like preprocessing, then the testing data set is ready to use for test or analysis the data. With the testing data set, depending on the dependent variable the provided output from the algorithm is representing the predicted splitting red circle countries.

In that case, when the polynomial regression algorithm is applied then we get another graphical output. But in the polynomial regression, no need to divide the data set like as training or testing set. But need to convert the categorical value to numerical value which is mandatory for this algorithm.

Output of the Polynomial Regression Algorithm (PRL),

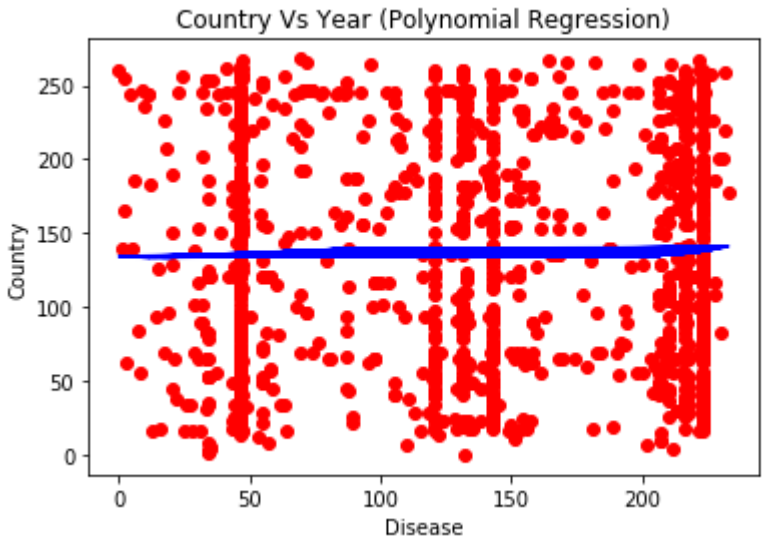

Figure 5: Polynomial regression line (PRA)

The above graphical representation is between disease and country where the year value is generated as the predicted value.

In this graph, the disease value and the country value is representing as the numerical value.

CHAPTER 5
CONCLUSION

5.1 Conclusion

Now time diminish the influence of disease, diseases prediction is most significant. This paper is developed using Linear Regression and Polynomial Regression. We made a prediction model using data mining and machine learning in this paper, to predict future diseases outbreak. Aim of our prediction model to make easier disease prediction process.

5.2 Future Scope

Now in this paper uses some data mining algorithm to predict disease. In future we will visualize the date of future disease outbreak predictive time. Also will develop a system, that every time analyze current data and produce predictive result and that system playing great rule to make future of human life better.

References:

[1] "Predictive Data Mining for Medical Diagnosis: An Overview," *International Journal of Computer Applications,* vol. 17, no. 8, p. 0975 – 8887, 2011.

[2] S. Palaniappan, "Intelligent Heart Disease Prediction System Using," *IJCSNS International Journal of Computer Science and Network Security,* vol. 8, no. 8, p. 343, 2008.

[3] M. F. M. Ghani, "Intelligent heart disease prediction system using data mining techniques," *Researchgate,* vol. 08, p. 08, 2018.

[4] "Data Mining and Knowledge Discovery in Databases: Implications fro scientific databases," pp. 2-1, 1997.

[5] P. C. S. Sellappan, "Model-based Healthcare Decision Support System", Proc. Of Int. Conf. on Information Technology in Asia CITA," *Researchgate,* vol. 05, pp. 45-50, 2005.

[6] S. S.Sudha, "Disease Prediction in Data Mining Technique – A Survey," *International Journal of Computer Applications & Information Technology,* vol. 2, no. 1, 2013.

[7] N. H. a. I. A. K. Ashfaq Ahmed K and Sultan Aljahdali, "Cancer Disease Prediction with Support Vector Machine and Random Forest Classification Techniques," *IEEE,* 2012.

[8] M. a. F. M. a. Fatima, "Survey of Machine Learning Algorithms for Disease Diagnostic," *Journal of Intelligent Learning Systems and Applications,* pp. 1-9, 2017.

[9] "Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach.," *International Journal of Scientific & Engineering Research,* vol. 4, pp. 680-685, 2013.

[10] G. Sathyadevi, "Application of CART Algorithm in Hepatitis Disease Diagnosis," *IEEE,* pp. 1283-1287, 2011.

[11] A. a. S. Sarwar, "Intelligent Naïve Bayes Approach to Diagnose Diabetes Type-2.," *ICNICT,* vol. 3, no. 14-16, 2012.