International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020
ISSN: 1475-7192

ENVIRONMENTAL ANALYSIS BASED ON CAUSES OF DEATH: A SURVEY

[1]M. Uma Devi, [2]Vishwanand Vyas, [3]Nikhil Agrawal

ABSTRACT--*Every environment has different factors based on which we can analyze the factors such as medical concerns, resource allocations, mortality factors and various others. Information is collected from the analysis of Causes of Death. This helps in analysis of death caused by lack of vaccination, environmental factors, and natural causes. By categorizing the data on the basis of age we can determine the major causes of death in a particular age group which in turn helps hospitals and doctors to prepare accordingly. It will help hospital and healthcare organisations to make better health care decisions and public policy makers to make better policies. A widely used method in the studies of healthcare is statistical method. Though these method are very useful in public health and clinical research policy making, these methods are not capable of finding relation among health condition on their own. Thus we aim to cope with the above challenge by giving input from different datasets and then producing an output which can give more accurate results and better understanding of the situation of the area.*

Keywords- *Data Mining, Bit Term TropicModel, Feature Selection, Classification &Regression Tree(C&RT), CoherenceScore*

I. INRODUCTION

Data mining is a method of retrieving useful and meaningful information from database corpus.

The outcome of the extracted data can be analyzed for the future planning and development perspectives. With the help of data mining techniques like regression, clustering, prediction and other we will be able to device a pattern and then analyses the given data set from the census.

More and more volume of data that is being generated every year makes retrieving useful data from that dataset very difficult. Data warehouse is used generally to store the information. It is a storage house for data which is gathered from many different sources which includes corporate databases, information from internal systems, and data from external sources. Analysis of data includes simple query and reporting, statistical analysis, more complex multidimensional analysis, and data mining.

Data Mining: It is processing data and analyzing it in different ways and then summarizing the information in a way that is useful and establishes relationship with data. There are two types of data mining: descriptive and predictive. Former gives information about pre-existing data and predictive mining is used to make predictions based on the dataset.

The five steps in decision-making can be identified as follows:

[1] *Computer Science and EngineeringSRMIST, Kattankalathur,Chennai, Tamil Nadu*

[2] *Computer Science and Engineering,SRMIST, Kattankalathur,Chennai, Tamil Nadu*

[3] *Computer Science and Engineering,SRMIST, Kattankalathur,Chennai, Tamil Nadu*

Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020

- Develop standard reports.
- Identify exceptions: unusual situations and outcomes that indicate potential problems or advantages.
- Identify causes of the exceptions.
- Develop models for possible alternatives.
- Track effectiveness.

II. LITERATURE SURVEY

There can be multiple causes of deaths and it opens a wide field to be analyzed. [8] **Mohammad Hossein Saraee** et al. In their paper they have worked on the deaths caused in children due to the accidents and have given suggestion on how the death rates due to accidents can be brought down. They have used techniques like Feature Selection, c&r tree algorithm, Blank Handling, Pruning, bayesian network to anayze the causes. The data mining method that is used are Bayes' theorem and decision tree. The author concluded that the rate of mortality in patients who are getting primary treatment before arriving to the emergency ward of the hospital, is way lesser than those who do not get any type of care before reaching the hospital, and this point focuses on value of pre-hospital care, necessity of ambulance and importance of learning the first aids.

[9] James K. Tamgno et al. Developed a mobile app to collect and store the data regarding the medical autopsies. The system will automate the collection, the sending and the insertion of the data in a database centralized. Verbal Autopsies is used for comparing local and national differences in mortality ratios, identifying main health problems, the surveying of trends over a period of time, and the growth of healthcare progrms.

[1] Hanyu Jiang, Hang Wu et al.in their paper have used techniques like Trend analysis, Topic model, Biterm topic model, Coherence Score to analyze the causes of deaths in united states of America between the years 1999-2014.The different topics are first clustered and then coherence score is used to value the quality of discovered topics. They have used the death certificates of people to collect the mortality data and thus is limited to the availability of death certificates. Their aim is to provide feedback of their analysis to clinical practitioner and public health policymaker to provide better health care service.

[3] Hesham Abdo Ahmed Aqlan et al. Have used web mining techniques to predict the possibility of deaths in future and particular diseases of interest. They used big-scale digitally stored histories which are captured for duration of 18 years from news reports of Queensland Government archive to make real-time predictions. The prediction of death is done using many different types classifiers and results are calculated using error as a metric. They have done modelling of data and and forecasting of events which is also known as time series analysis and time series forecasting. They have proposed death prediction and analysis of deaths in this paper. They have used four separate ways to predict deaths from large scale data.

[4] Ogochukwu C.Okeke et al have done census analysis to analyze geo spatial data distribution for Nigeria which is a very populous country. Their effort is dedicated to harness the power of data mining techniques to develop a data mining model which is apply able to and can be used for the analysis of census data. They aim to provide government with the intelligence for strategic planning, tactical decision-making, better policy formulation and for better-informed business decisions. They have used decision tree for doing this analysis. The techniques used were decision tree algorithm, structured system analysis and design methodology. With the proposed system

their aim is to provide better facilities to all the citizens of Nigeria who might not be getting the proper facilities and exposure.

[6] Munaza Ramzan in his paper has done classification and characterization for the treatment strategies of critical diseases such as cardiovascular diseases, cancer and diabetes which are major causes of deaths worldwide. He has used data mining techniques such as weka to do the analysis of the medical data. Some of the other techniques used are J48, naive-bayes and random forest. His work calculates the disease categorization using 3 types of machine learning algorithms using WEKA Tool. His works shows that Random forest is the best classifier for disease categorization because it runs very effectively on large data sets.

[5] Du Zhang et al. Have done mining of vital statistics data on causes of deaths in California state. They have used data mining tool which is known as Cubist which they used for building a predictive model using dataset which consists a period of over nine-years and more than two million cases. They have used techniques like committee model, cubist model and data selection strategies to do the analysis. The aim of their study is to discover information that can be used for obtaining knowledge regarding many different aspects of mortality in the state of California, to offer aid for decision-making process, for predicting health issues related to the causes of death, and provides customers with useful information services.

Paper	**Author**	**Techniques**	**Work done and Advantages**	**Disadvantage**
Data Mininig in Medical Data with child mortality	[8]Mohammad Hossein Saree & Bahare Zibanezahad	Bayes Theorem & C&RT algorithm Bayesian network	They have worked on the deaths caused in children due to the accidents and have given suggestion on how the death rates due to accidents can be brought down. Helps in finding the death causes in children and reducing it.	Limited to only children and not for all age groups.
Analyzing relation between medical conditions using data mining	Rachel L freeman & Laura A Riffel	LERS(learning from examples using Rough Sets) & Rule Induction	Analyzes the causes of death due to diseases. Sequential death sets were used with snapshots of patients behavior and health record. Techniques like	Limited only to the diseases and doesn't account for accidents and suicides.

			LERS and rule induction were used.	
Weka Classification of medical data	[6] Munaza Ramzan	Classification algorithms (naïve bayes) , random forest & WEKA Tools	The characterization and classification of the treatment strategies for critical diseases like cardiovascular diseases, cancer, and diabetes are done. He has used data mining techniques such as weka to do the analysis of the medical data Analyzes the treatment pattern for critical diseases	Limited to only critical diseases like cancer.
Verbal Autopsies, Mobile Data Collection for Monitoring and Warning Causes of Deaths	[9] James K. Tamgno, Roger M.Faye,	Verbal autopsies.	Developed a mobile app to collect and store the data regarding the medical autopsies.The system will automate the collection, the sending and the insertion of the data in a data base centralized. Verbal Autopsies is used for the surveying of trends over time, comparing local and national differences in mortality ratios, identifying serious health problems, and the growth of	

			healthcare programs.	
Causes of Death in the United States, 1999 to 2014	[1] Hanyu Jiang, Hang Wu, May Dongmei Wang,	Topic model, Biterm topic model, Coherence score, clustering.	Techniques like Trend analysis, Topic model, Biterm topic model, Coherence Score were used to analyze the causes of deaths in united states of America between the years 1999-2014. Coherence score was used to measure the quality of the topic.	Limited to the availability of the death ceritficates.
Death Prediction and Analysis Using Web Mining Techniques	[3] Hesham Abdo Ahmed Aqlan, Shoiab Ahmed, Ajit Danti	Web mining techniques.	Have used web mining techniques to predict about the possibility of death in upcoming time and particular diseases of interest. Prediction of death is done with the help of many different calculated using error as a metric. They have done modelling of data and forecasting of events which is also known as time series analysis and time series forecasting. They have proposed death prediction and	Based on the news reports which can be incorrect at times.

			analysis of deaths in this paper	
Using Data-Mining Technique for Census Analysis to Give Geo-Spatial Distribution of Nigeria	Ogochukwu C.Okeke and Boniface C, Ekechukwu	Decision tree algorithm, Structured System Analysis, Design Methodology	Data-mining technique is used to develop a mining model which can be used for the analysis of census data. They aim to provide , better policy formulation and for With the proposed system their aim is to provide better facilities to all the citizens of Nigeria who might not be getting the proper facilities and exposure	

III. TECHNIQUES USED

In order to perform the present system for analysis of the statistics, knowledge in the following fields is required.

1.Data Mining

2.Statastics

3.Python or R programming

4.Data Analysis Tool

Tools to represent the statistical data in the form of graphical diagrams for better understanding and prediction.

3.1.DATA MINING

Knowledge in the field of data mining is crucial for working the project as it involves various techniques such as regression, Classification and Clustering.

Regression is a method which is used for modeling and analyzing the relationships between variables, the way they contribute and are related to producing a particular outcome altogether. A linear regression is a regression model that comprises of linear variables.Single Variable Linear Regression is a method that is used for modeling the relationship between a single input independent variable (feature variable) and an output dependent variable using a linear model.

Classification is a method that makes a collection of items in order to target classes or categories. The aim of classification is to precisely and correctly predict the targeted class for every case in database. Classification are discrete and do not have any order. Floating-point, Continuous values are used for indicating a numerical, instead of a categorical, target. A predictive model with a numerical target uses a regression algorithm, not a classification algorithm.

Clustering or Cluster analysis is the process of grouping a set of objects in a manner that objects in the same group (know as cluster) are more similar (in some ways) to each other rather than those in other groups (clusters).

3.2 STATASTICS:

In the data set the mathematical technique such as topic models, average estimation and segregation all require the knowledge in the field of statistics. Mathematical knowledge is required to formulate and standardize the data set into different classes based on which the data mining algorithms are carried out.

3.3 Python & R programming

Python and R are both open-source programming languages having a large community. R is generally used for the purpose of statistical analysis while Python is used for providing a more general approach towards data science. New libraries or tools are added regularly to their respective catalog.

R and Python are best and most used programming languages in terms of programming language that are dedicated for data science. Python and R requires a lot of time to master them.

Python is a scripting language known for vast compatibility with other useful tools such as Anaconda ,Juypter and pandas.

3.4 DATA ANALYSIS TOOLS

Weka tool: Weka is a machine learning tool which is very easy to learn. Its interface is intuitive and helps to get the job done quickly. It provides options for clustering, classification, visualization, regression, data pre-processing and association rules. Most of the steps that are used for model building are achieve able using Weka. It is built on Java programming language.

Datawrapper Tool: Datawrapper an be used to make interactive charts. It is an online data-visualization tool. It is very easy to use and produces effective graphics. The data can be either uploaded from CSV/PDF/Excel file or can be pasted directly into the field. It will generate a visualization such as a bar, map, line. Graphs that are developed through the datawrapper can be embedded into any website with ready-to-use embed codes. So many reporters and news organizations use Datawrapper to embed live charts into their articles.

IV. PROJECT DESCRIPTION

This project is done to Analyse the environment based on the causes of deaths in an area. There can be numerous causes of deaths in an area and thus the process of data mining is used to find the major causes of deaths. Data set of a developing country which is also very populous, is used. In very populous regions due to high death rates the causes of death sometimes go unnoticed until its too big. This project is going to help identify the cause a lot earlier. The project is based on analyzing the data set and formulating a model using techniques of

classification, clustering, regression. For understanding the dataset, it is represented in a graphical method by using the visualization tools like WEKA. Initially when a data is collected it may have many redundant data and improper entries. Feature selection algorithm is the technique which is used for removing the redundant data. Classification technique is used to classify the data and then clustering technique to cluster the data into different data sets (Trend analysis). Once the data is segregated based on the different death causes then it will be represented using bar graph and pie charts. The different topics, which are clustered together, will be analyzed using coherence score. Coherence score will give the quality of the topics i.e. Its factor of influence can be determined using coherence score. The representation of data in graphical way will show the major factors of death and methods canbe suggested to bring down the deaths due to the major causes. The aim of this project is to increase the standard of living in the highly populated areas by reducing the death rates and providing better facilities.

4.1 BENEFITS OF OUR PROJECT

1. The analysis will provide results based on which suggestions can be made to the government to improve the management and services in terms of medical supplies provided to the affected area.

2. It will help to understand the major factors responsible for major diseases and death and their effects in the area.

3. This analysis will also help to reflect upon the crime rates and deaths due to accidents in the area and thus immediate authorities can be informed about the same to bring down the crime rate and to improve the traffic system in the area.

4. Upon taking necessary action on the problem the results of the steps taken can be seen in the next survey.

4.2 FUTURE SCOPE

The analysis will provide loopholes in the present system and upon rectification of the problem focus can be on the overall development in the medical standards of the area making the environment healthy. This will later increase the economic value of the area and the residents of the area will gain benefits such as increased property values and decreased medical diseases and rodent infestation. Not only will the project help in economic way but will also increase the area's environmental value by decreasing the contaminants and the disease spreading animals and insect.

From the management point of view it will be efficient for distributing the medical supply and needs to the areas based on the results rather than provide some with more supplies and other with less supplies. From the result analysis the need for the medical supply and specifically for which diseases will be explained and provided as the result of this project.

4.3. CHALLENGES

The dataset available is not always reliable and may contain certain contaminations. The method to collect the data is not efficient in many countries and finding the relations between different datasets may sometimes be difficult.

Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020

In the near future with the development of digital economy and increase in storing data digitally in large scale the process of gathering data will become efficient and will prove useful. The field of data mining is also increasing daily with new classification and analyzing techniques.

V. CONCLUSION

In this report, study was done on how to extract and analyze the important information from the data set. Big datasets were represented into a graphical representation by using certain influence factors procured by analyzing the data set. Based on the representation and the derivation from the graphical representation, necessary step can be taken to ensure that the efficiency of the system is increased and the target objective is obtained with some room for improvements and future upgrades. The proposed model of analysis tends to be improvised in the future and has great scope for the upcoming technological advancements with everything becoming digitalized and increasing the ease with which management of the area can be increased

REFERENCES

1. Hanyu Jiang, Hang Wu, Student *Member*, and May Dongmei Wang, Ph.D., on"Causes of Death in the United States, 1999 to 2017"in *IEEE,*2017
2. A. Jemal, E. Ward, Y. Hao, and M.
3. J. Thun, "Trends in the leading causes of death in the United States, 1970-2002," JAMA,vol. 294, no.10, pp. 1255–1259, 2005.
4. Hesham Abdo Ahmed Aqlan, Shoiab Ahmed and Ajit Danti on "Death Prediction and Analysis Using Web Mining Techniques*"* published on 2017 International Conference on Advanced Computing and Communication Systems (*ICACCS* -2015), Jan. 06 – 07, 2017, Coimbatore, INDIA
5. Gerami Farzad, Bartashak
6. Masoumeh, Kourosh Rocky,Razieh Honarmand,"Prediction of Workplace Accidents with Knowledge Discovery
7. Approach Using Weka software,"Nova Explore Publications, Nova Journal of Engineering and Applied Sciences, Vol. 2(5), May2014.
8. Du Zhang, Quoc Luan Ha and Meiliu Lu on "Mining California Vital Statistics Data" published in 2001 IEEE
9. Munaza Ramzan on "Comparing and Evaluating the Performance of WEKA Classifiers on Critical Diseases" published on 2016 IEEE
10. Mr. Chintan Shah, Dr. Anjali G. Jivani, on "Comparison of Data Mining Classification Algorithms for health analysis",IEEE,2013
11. Mohammad Hossein Saraee Advanced Database Systems, Data Mining and Bioinformatics Research Laboratory on " Applying Data Mining In Medical Data With focus on mortality related to accident in children" in IEEE 2008
12. James K. Tamgno, Roger M. Faye, Claude Lishou on "Verbal Autopsies, Mobile Data Collection for Monitoring and Warnin Causes of Deaths" in ICACT 2013.