ISSN No:-2456-2165

Extraction and Processing of Medical Data using Data Mining Techniques

ADIKWU JOSEPH O.
Department of Mathematics/Statistics,
Benue State Polytechnic, Ugbokolo, Benue state, Nigeria.

Abstract:- This work is on the use of data extraction and processing techniques in extracting data from medical dataset for analysis was done using data mining technique and provides useful tips and procedures for an easier process of obtaining needed data from large dataset such as medical records. Statistical analysis using ANOVA was done at the 5% significance level which indicated a high significance in the infection level for different states investigated. The system of extraction has common problems (challenges) and handlings in the various states were carefully studied for the dataset. Findings will be of great assist to health organizations in making informed and concise decisions. Recommendations were proffered which include among others, the following, that the fact that existing classification of medical data should be improved upon.

Keywords:- *Medical data, Extraction, Processing, Mining, Dataset.*

I. INTRODUCTION

A. Background of the study

In our today's world that is endowed with rich information in form of data (large and small), data extraction and processing has come to be of immense benefit. It is worthy of note to state that this knowledge of the availability of data is quite comforting. One major challenge is the enormous volume of the available data for researchers. When more information are available, it takes longer to find useful insights into needed results to achieve intended aims and objective of an intended investigation.

That is why this work is on data extraction and processing of data using data mining techniques. Data mining has been defined as the process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems (Data Mining Curriculum, 2006). The research explored majority of the aspects of data extraction and processing, including what it means, its stages, the techniques involved as it is relevant to analysis of medical data which in most cases presents very large set.

To have a clear understanding and view of the application/use of data-mining procedure and techniques to medical big data, this research introduces the concept of databases from public medical facilities and summarizes same for use in medical and related researches. This research also illustrate data-mining algorithms that are commonly used in medical practice beside some specific cases to serve as a veritable tool to clinical researchers have a clear and detailed understanding of the application and uses of data-mining techniques to medical data of large volume.

According to Penz et al (2007), Data Analysis is a tool for extracting the jewel of truth from the slurry of data. "And data extraction and processing and statistics are fields that work towards this goal. While they may overlap, they are two very different techniques that require different skills.

Data-mining has two kinds of models: descriptive and predictive. Predictive models are used to predict unknown or future values of other variables of interest, whereas descriptive models are often used to find patterns that describe data that can be interpreted by humans (Kantardzic et al; 2003)

Statistics, as an act of science, form the nucleus of data extraction and processing, which encompasses the entire procedure of analysis of data. Statistics is an aspect of data extraction and processing that provides the tools and analytical techniques for dealing with large volumes or amounts of dataset. Statistics helps to identify model that further assist in identifying differences between random noise and significant findings - providing a theory for estimating probabilities of predictions and more. As a result, we can say that, data extraction and processing and statistics, as an act or method of data-analysis, is of great help that enhances better decision - making.

It is the science of learning from data which includes but, not limited to collecting and organizing to analyzing and presenting data. Statistics is an act of science that focuses on probabilistic models, which draws conclusion as specific inference, using data.

II. STATEMENT OF PROBLEM

Some of the problems of data extraction and processing that this research addressed are aimed at solving include:

- Mostdata analysis tools are always complex and difficult to use. This research will help make it easier to classify data.
- In other to handle the act and use of data extraction and processing techniques, we have to find the most suitable strategy. In order to do that, we have to detect the problem type. Usually, data extraction and processing project involves a combination of different problem types which together solve the problem.
- Data extraction and processingdemand considerably significant amount of dataset or bases, making the process difficult to handle.

ISSN No:-2456-2165

III. AIM OF THE RESEARCH

This research will help design classifiers to handle high dimensional classification problem and provide useful tips and procedures for an easier process of obtaining needed data from large dataset as medical records.

IV. OBJECTIVE OF THE RESEARCH

Data extraction and processing converts information into usable knowledge. It has brought to bear a world of possibilities for business and business decision making. This field of computational statistics compares vast amount of medical and similar data and is used by researchers and interested parties to study and predict the behavior of the users of medical facilities as either patients, customers or consumers of services. The research will help hospital and other similar business setups to know the necessary precautions and some necessary changes that needed to be done in their daily operations to improve efficiency rate and quality of services given to outpatients.

V. BENEFIT OF THE RESEARCH

The act of data extraction and processing avails researchers, processes and procedures for solving problems in this information and computer technology age. Some of the benefits of this research are; it will help researchers and interested parties gather reliable information, provide useful tips and procedures for an easier process of obtaining needed data from large dataset and help businesses make informed decisions.

Data extraction and processing models assist businesses and companies acquire well informed and processed information, thereby increasing the company's profit margin through adjustments in processes and operations.

VI. WHAT IS DATA EXTRACTION AND PROCESSING?

Putting the words of data scientists Lussier, et al (2001) in another way, data extraction and processing is "the tasking processes of identifying acceptable, new and original, powerfully useful and most importantlyunderstandable trends in data".

Modern day technological know – how has boosted the process of extraction of obscured predictive information from large databases, along with a confluence of various other frontiers or fields like statistics, data visualization, database management, artificial intelligence, pattern recognition and machine learning.

Data extraction and processing can be described as the technique of analyzing large amounts of data and datasets, bringing out important intelligence that can assist businesses (in his case, medical facilities) handle challenges, predict trends, mitigate risks, and find new opportunities. Data extraction and processing is like actual extraction and processing because, in both cases, the processors are searching through large volumes of informationsources to obtain important and materials.

Also, the act of data extraction and processing involves proving a link and obtaining models, anomalies, and correlations to handle challenges, developing actionable information in the process. This area of study covers a large and different technique that involves a lot of varying partsof whichsome are mistaken to be data extraction and processing.

With data extraction and processing, an individual applies various methods of statistics, data analysis, and machine learning to explore and analyze large data sets, to extract new and useful information that will benefit the owner of these data.

By using data extraction and processing, an organization may discover actionable insights from their existing data. For example, by analyzing social media posts, a snack foods company may be surprised to learn that their largest market is single dads.

VII. STEPS INVOLVED IN DATA EXTRACTION AND PROCESSING

If a researcher wishes to undertake a data extraction and processing project, there are some basic steps needed to follow; such steps are outlined below:

A. Understand the intention

What does the researcher have in mind, the objective of the research and what will determine the acceptability of the process followed?

B. Understand the Data

Discover the type of numerical or non – numerical information that is required to obtain the solution to the challenge after which the data is collected from the relevant or appropriate source.

C. Data presentation

Sort out the data to avoid duplication, corrupted or missing data and present the obtained data properly in a way it can be used to solve the related challenges.

D. Model the Data

Use step – by – step (algorithm) method to determine data patterns. Then create, test and evaluate the model.

E. Evaluate the Data

After obtaining the data, appropriate analysis procedure is employed to analyze the obtained data.Take a decision as to how relevant results or outputs obtained by the model will assist in meeting the goal or remedy the challenge.

F. Apply the Solution

Release whatever outcome (results) from the research findingto the people in charge of making decisions.

ISSN No:-2456-2165

VIII. REVIEW OF RELATED LITERATURES

Data extraction and processing has come to be generally accepted round the world to be a very important study area, lying at the point where statistics, data management,machine learning, artificial intelligence,pattern recognition and the likes have a common connotation.

All of these are shown to be related with certain aspects of data analysis, so they have much in common; but each also has its own distinct flavor, emphasizing particular problems and types of solution (Sager, et al., 2008).The astronomical growth and conglomeration of information technologies, digital networks, software and database systems and the availability of enormous amount of electronic data provide researchers with vast new resources that can be analyzed to optimize managerial decision, uncover financially valuable patterns, minimize investment risks, make successful strategic decisions, and so on (Costea&Eklund, 2009).

Since the introduction of computers and the information age, statistical challenges have grown exponentially in magnitude and complexity. Challenges in the areas of data storage, organization and searching have led to this new area of "data extraction and processing"; statistical and computational problems in biology and medicine have created "bioinformatics." Vast amounts of data are being generated in many fields, and the statistician's job is to make sense of it all: to extract important patterns and trends, and understand "what the data says." This is generally described as "learning from data". In other words, data extraction and processing refers to the search of large, high-dimensional, multi-type data sets, especially those with elaborate dependence structures or patterns where the search for valuable structure or patterns is based on statistical methodologies (Hastie, et al., 2005).

Even, some interesting studies have been applied in fraud detection in several business area including medical records and systems (Ortega, et al., 2006). In order to study significantly large volume of data analysis research, investigators (researchers) and users of research report, have adopted established step – by – step process from statistics, and other studies and have also advanced a new approach aimed at large data extraction and processing problems (Hand, et al., 2001). Also, data extraction and processing can be defined as an automatic or semiautomatic patterns discovery in great amounts of data, where these patterns can be perceived as useful (Witten, 2011).

IX. MATERIALS AND METHODS

A. Data Collection and Preparation

The data used for this research work are collected from hospitals' outpatient clinic records from selected public and private hospitals in the North Central states of Nigeria which include Benue, Nassarawa, Kogi, Niger, Kwara, Plateau and FCT.

The dataset size are relatively large in volume for the data mining process, before analyzing the data, a Simple Random Sampling method was used in drawing samples of the data. Simple Statistical tools was used for the analysis.

The samplesize that was chosen for this research is according to the following:

The original dataset size: N

Confidence interval (accepted margin of error): α – value of 5%

Confidence level: 95%

Table 1: Number of persons infected by diseases

States	Diseases				
	Covid – 19	**Malaria**	**Typhoid**	**Tuberculosis**	**Lassa Fever**
Benue	2129	94386	360	38	8
Nassarawa	2720	71326	233	17	3
Kogi	5	64291	395	10	3
Niger	1148	115086	212	12	2
Kwara	4601	93764	522	18	2
Plateau	10252	59117	305	75	9
FCT	28618	32883	161	16	3

Source: Nigeria Centre for Disease Control office, Abuja.

The report shows that in Nigeria, as at January, 2022, two hundred and eleven cases of Lassa fever were confirmed in laboratory test conducted, and fatality of 40 (19%). These cases were reported from 14 states. Some of which are captured in the table above.

Table 2: MINITAB outlook of One-way ANOVA for the diseases

Source	DF	SS	MS	F	P
Factor	4	30906065371	7726516343	45.51	0.000
Error	30	5093143789	169771460		
Total	34	35999209161			

S = 13030 R-Sq = 85.85% R-Sq(adj) = 83.97%

ISSN No:-2456-2165

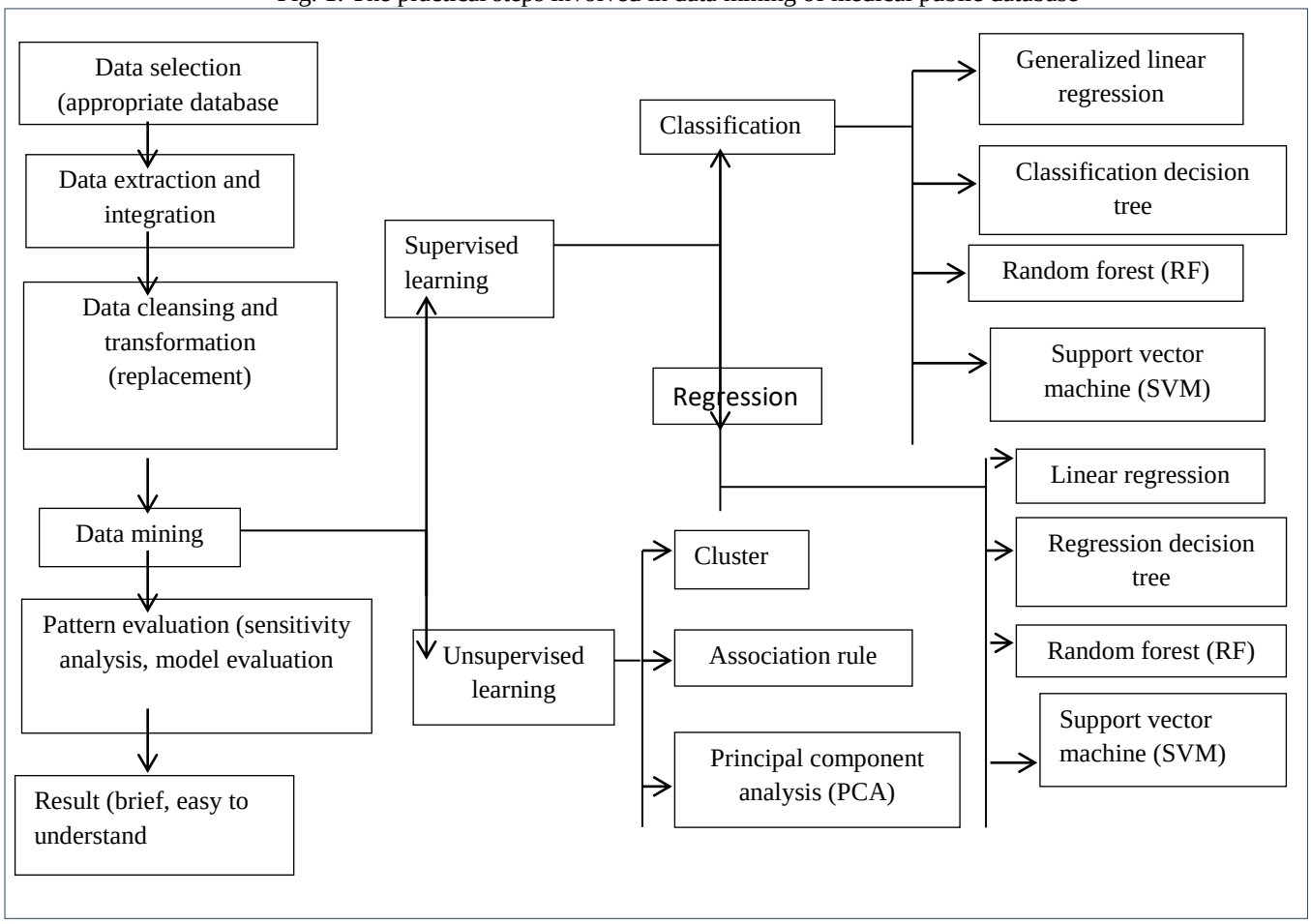
Table 3: MINITAB outlook of One-way ANOVA for the diseases in states (Pooled)

Individual 95% CIs For Mean Based on
Pooled StDev

```
Level        N   Mean  StDev  ----+---------+---------+---------+-----
Covid – 19   7   7068  10074    (---*---)
Malaria      7  75836  27338                        (---*---)
Typhoid      7    313    124  (---*---)
Tuberculosis 7     27     23  (---*---)
Lassa Fever  7      4      3  (---*---)
 ----+---------+---------+---------+-----
0     25000     50000     75000

Pooled StDev = 13030
```

Fig. 1: The practical steps involved in data mining of medical public database



X. RESULT

The research revealed some very pertinent and necessary precautions that need to be adhere to in management of records. It also shows some important changes that need to be done in the day – to – day operation of businesses such as medical facilities that has to do with record keeping especially as it relates to medical records. Data mining sometimes can be mistakenly used wrongly, resulting in an outcome that looks significant which can really tell or predict the behaviour of data in the future and cannot be duplicated (produced a second time) andjuxtaposed side – by – side any new sample of similar data, which means very small use of it.

XI. CONCLUSION

The research work when properly put into use, will help accelerate the rate at which informed decisions will be made and also improve the knowledge of what is relevant and also make good use of information to obtain likely results. This work has also highlighted some limitations of data mining processes and discussed certain further research concerns.

Using of large dataset has brought about changes in most all aspects of today living, with the use of large dataset combined with data-mining techniques which can enhance the status quo. One of the objective of this research is to help medical researchers in having a better knowledge of the uses and use of data-mining tools and skills on medical large dataset and public medical databases to further their project objective in order to benefit medical personnel and their patients.

REFERENCES

[1.] Sager N., Friedman C. & Lyman M. (1986): The analysis and processing of clinical narrative. In: Salamon R, Blum B, Jorgensen M, eds. Proceedings of the fifth conference on medical informatics. Washington DC, USA: Elsevier Science Publishers; B.V,
[2.] Lussier Y., Shagina L. & Friedman C. (2001): Automating SNOMED coding using medical language understanding: a feasibility study. ProcAMIASymp.
[3.] Baud R. (2004): A natural language based search engine for ICD10 diagnosis encoding. Med Arh
[4.] Penz J., Wilcox A. & Hurdle J. (2007): Automated identification of adverse events related to central venous catheters. J Biomed Inform
[5.] Costea M., Eklund W &Aronsky D, (2000): Automatic detection of acute bacterial pneumonia from chest x-ray reports. J Am Med Inform Association.
[6.] Witten I. H.(2011): Data Extraction and processing Practical Machine Learning Tools and Techniques,3rdEdition. ElsevierInc., USA.
[7.] Hastie E. &Peissig P. (2005): Study of effect of drug lexicons on medication extraction from electronic medical records. Pac SympBiocomput
[8.] Ortega A. & Martinez C. (2006): An algorithm to derive a numerical daily dose from unstructured text dosage instructions. Pharmacoepidemiol Drug Saf
[9.] Kantardzic M. Data Mining: concepts, models, methods, and algorithms. Technometrics. 2003;45(3):277.
[10.] "Data Mining Curriculum". ACM SIGKDD. 2006-04-30. Retrieved 2022-10-2.
[11.] Sager, N; Chi, E; Friedman, C (1986). The analysis and processing of clinical narrative. Medinfo; Elsevier.
[12.] Soldaini, L; Yates, A; Yom-Tov, E; Frieder, O; Goharian,N (2016). Enhancing web search in the medical domain via query clarification. Information Retrieval Journal.doi:10.1007/s10791-015-9258-y.
[13.] Uzuner, O; Luo, Y; Szolovits, P (2007). Evaluating the state-of-the-art in automatic deidentification. JAMIA 2007.
[14.] World Health Organization (2010). International statistical classification of diseases and related health problems 10th revision, edition 2010.Geneva, Switzerland.
[15.] Zheng, J; Yu, H (2016). Methods for linking EHR notes to education materials. Information RetrievalJournal.doi:10.1007/s10791-015-9263-1.