International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (online): 2347-5552, Volume-11, Issue-3, May 2023
https://doi.org/10.55524/ijircst.2023.11.3.6
Article id- IRP1380, Pages 32-37
www.ijircst.org

Techniques for Data Mining Prediction in the Health Care Sector

Aditya Kumar Tripathi[1], and Anu Sharma[2]

[1,2] Assistant Professor, Department of Computer Science & Engineering, Teerthanker Mahaveer University, Moradabad, India

Correspondence should be addressed to Aditya Kumar Tripathi; 1aditya250790@gmail.com

Copyright © 2023 Made Aditya Kumar Tripathi at al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT- Data mining is another term for knowledge discovery in databases (KDD). It's an interdisciplinary field that focuses on rooting meaningful knowledge from data in all sectors similar as health, education, and business. Currently, with the covid epidemic affecting everyone and rising coronavirus cases causing nursing home beds, oxygen, vaccines and individuals to be denied by hospitals, the health structure of the elderly is in the spotlight. There's a wealth of information accessible in the medical world about these diseases. Data booby-trapping concepts may be used to prize meaningful styles from this type of material in order to prognosticate unborn followings. This study emphasizes on several mining approaches that will be applied in the therapy assiduity to achieve the stylish results.

KEYWORDS- Mining, Classification, Clustering, Machine Learning.

I. INTRODUCTION

Researchers in health care are swamped in data yet famished for information. There is a multitude of health data as we know it, but it is currently lacking the analytical tools necessary to build correlations between them or reflect undiscovered patterns. When dealing with very large amounts of data, the traditional approach is extremely time-consuming as well as unusable. Mining is one of the greatest approaches that has received positive feedback from the government, healthcare groups, businesses, and other institutions. These strategies are mostly utilised for understanding large data and assisting hospital staff in providing better treatments and Doctors can use these strategies to browse through patient data and draw conclusions for procedures, OPDs, and other tasks. Knowing the patient's future based on data recorded electronically is one of the best qualities that may be employed in healthcare to improve patient wellbeing.

Data mining has absolutely been a part of our lives since before the dawn of data science and its ventures, if we include the medical industry in that, there is no exception. After the pandemic covid, the internet can be used to successfully implement health objectives, for example, people can use the internet for various medical purposes in different age groups.

Data mining is rapidly gaining ground all over the world. Everyone are presently hesitate of use contemporary healthcare services such as immunisation. Uploading the data to a website aids the authorities in identifying flaws. Countries such as the United States, the United Kingdom, and the European Union have advanced substantially because all government programmes, projects, and economic growth are now available online. Making information and government initiatives available online promotes transparency, good governance, and improved public-government relations. Data mining techniques can be used to successfully address medical problems in a country with limited resources and a high population density. The lack of expert doctors in remote places can be best addressed by building an online medical consultation app in the Play Store that provides better health care facilities to the general people.

Transparency, information, good governance and public-government relations can be improved by making research works and their conclusions available online. Data mining techniques can be successfully used to solve medical problems in countries with limited resources and high population density. The shortage of specialist doctors in remote areas can be best overcome by creating an online medical consultation app in play store and providing better healthcare facilities to the common people.

The process of discovering information in a system is called knowledge discovery. Big data can be used to find information. Data mining uses 7 phase to extract information from large amounts of data. This includes various methods including clustering, classification, summarization, association, ANOVA, and visualization. This article focuses on data mining methods that can be used in the healthcare domain. [9] This study also includes a brief review of the literature and a discussion of treatment strategies for various diseases. Our nation's medical institutions are a mess, and data-mining techniques help with regulation by recording vast amounts of patient data. Traditional methods are neither accurate nor simple and require current technologies such as data mining for optimal treatment and daily patient monitoring. Currently, big companies are trying to invest in healthcare artificial intelligence, building online counseling apps in the Google Play store, dealing with people in remote areas by expanding emergency services and proposing medicines through internet sites or applications based on ML algorithms. They compute the data make prediction based system.

A. Ground reality and existing issues in healthcare

Because commonly people are unaware of it, the issue is determining how people can benefit from it. The primary goal of this research is to deliver better amenities at a lower cost. Inadequate healthcare personnel: The current circumstance demonstrates how the crisis of medicators or teaching professionals happens the past Corona pandemic. During 2019, a report of survey revealed that India has approximately 250 health workers per 100,000 population in order to handle the issues that are expected in the near future and prepare to battle them.

II. FINDING USEFUL INFORMATION AND DATA MINING METHOD-

A. Technique of Data Mining

In general, data mining consists of seven steps:

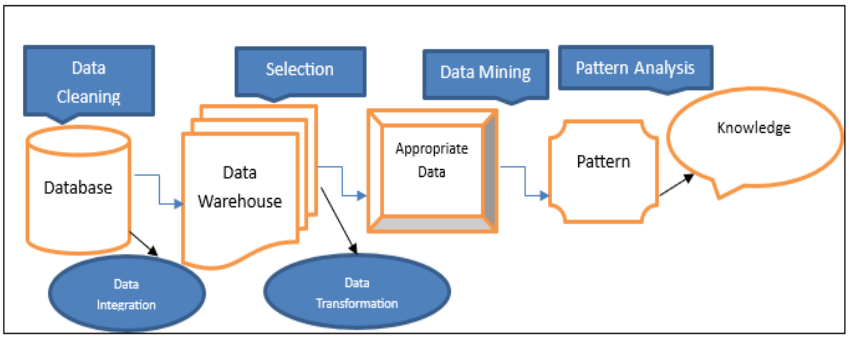

Figure 1: Data Mining Process

And all the steps stating from the data cleaning then integrations and further on accomplish using techniques used in different types of algorithms selected for a specific method of query. Each and every step of Figure 1 are given below with description.

a) Data cleaning

Generally, tables are collections of data groups, i.e. data kept in table's rows and columns to manage that it can be fluently penetrated, maintained, and streamlined. Row data drawing regulating of removing splutter and inconsistencies from data. drawing is performed as well to descry grammatical crimes, duplication, and missing data. For illustration, parsing the data to determine whether the supplied argument as statement respectable under the data standard.

b) Technique of Data Integration

According to this technique, data sources coming from different databases are gathered at one place. The data used is a mixture of diverse and homogeneous types that are kept in a storehouse.

c) Method of Data selection-

Data selection mean to select specific data utilizes for the analysis collect from the databases.

d) Method of Data transformation

Transform this data into a shape suitable for mining by performing rollup or aggregation operations.

e) Method of Data Mining

An important process in which ingenious methods are applied to arrangements to eradicate data styles.

f) Method of Pattern evaluation

Identify truly fascinating designs for describing knowledge in terms of some fascinating expedients.

g) Method Used to Finding Knowledge

Where visualization of outcome and knowledge exploring systems implemented to provide booby-trapped information to users. So, to summarize, KDD is a three-step process, namely

Preprocessing — Mining — Postprocessing In the preprocessing approach, data levels are refined through data culling, data integration, and data transformation.

It's one of the most caviling way in data mining process which deals with the remedy and transfiguration of the original dataset and again in Post reprocessing strategy is to uproot the knowledge in similar form that user can perception into data for better decision making and therefore it includes Pattern evaluation and visualization for pulling the knowledge.

III. METHODOLOGY

Baek et al. [1] Describing risk of depression across four platforms as effective, poor, severe, and truly unsafe, depending on the environment modified by multiple regression. It uses 14 variables to derive depression risk as layers accumulate as the data grows. This question can be answered in a short period of time through personal accomplishment. The prognostic values in this study are in unit intervals, between 0 and 1. Khan et al. [2] present a color data mining approach for diabetes detection, grouping, and vaccination. The conclusion drawn in this study is that in order to obtain accurate results, the data should be preprocessed and similar models should be used instead of one. Deren et al.

[3] Three methods including artificial neural network, decision tree, and logistic regression for bone cancer survival were compared with the predictive vulnerability of the three models. This study adopted statistical methods and data mining methods as exploratory methods. He developed predictive models and explored associations between independent variables and cancer survival.

These cases are used as datasets and 10-fold cross-validation for relative wait efficiencies of different data mining styles. The results of this study show that decision trees are better fashion predictors than others.

IV. HOSPITALITY AND SERVICES WITH DATA MINING

B Krishnaiah V [6] Care for a person's physical well-being is known as health service. The WHO states that providing first aid, which is everyone's fundamental need, as well as providing healthcare in remote locations are the goals of a healthcare system. Raising capabilities or fulfilling the demands of healthcare providers is another important goal of the healthcare system. The fair underpinning installations, such as the plutocrats paid by the populace to provide medical equipment, installations, etc., or the revenue accrued by state governments or medical insurance funds, etc., must be insured. There is now a vast amount of data in healthcare diligence, including case records, structures, and all health coffers, including the outfit, colourful complaint opinions, and lab tests.

Mahmoud H. [7] That might lead to issues including a lack of transparency, excessively important gratuitous care, disregard for cases, an overabundance of open cases, and many more. As people are now in the same state owing to COVID-19 pandemic, any of these issues may be decreased with the aid of these methods. Because forewarnings may not prevent hospitals from deteriorating, vaticination in health care is an essential responsibility, and data mining methods assist in finding inactive linkages in the data. The major goal is to raise the standard of living for better outcomes and less expensive therapies. [13] With the end of this, healthcare facilities may also be accessible in remote locations, and the number of premature deaths may decline. With analytical skills, we can explain and explain large data trends associated with improved response rates from patients to aid with the selection of remedies for ideal case situations. [4] A reputable medical firm that concentrates on hospitals and conferences grows its business and enhances its return on investment. Patient records may be easily accessed and well-treated with the analysis of data fraud decreasing. Since the information in books is only for show and serves no purpose, data booby-trapping allows for the diagnosis of problems and the recommendation of appropriate treatments, giving those who are physically unable to attend an edge. Several medical supply firms developed their goods by analysing how they will be used in medical settings, such as conferences.

adhering to the standards established for customised and altered treatment regimens. With the development of healthcare instruments, attentive nurses are better able to maintain a close eye on individuals and are also in charge of day-to-day monitoring. One of the main advantages of employing data extraction technologies within the medical that makes up one of the natural remedies for recovery. [15] investigates realistic ways to identify fake news on digital platforms in this context automatically in health care system. To begin, a massive number of current and correlated works were surveyed in an attempt to incorporate all possible features for detecting fake news, followed by exploratory data analysis to identify sources that frequently publish fake news and determine the most frequently occurring words in the title and body of fake and genuine news.

A. *Hierarchical presentation of information Mining*

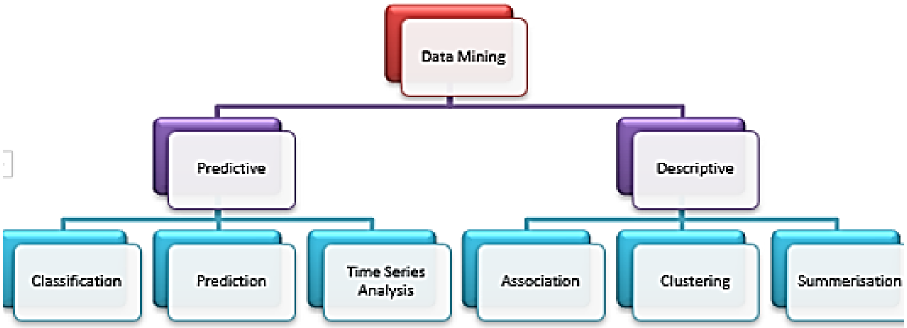

Figure 2: Hierarchical presentation of Mining

In this Figure 2 Predictive method and Descriptive method and both types further scattered in 3 parts to continuing in hierarchy. Its explanation are given below.

a) Predictive Method

This model focuses on cast of the existent like credit threat etc. It's the most common habituated system as grounded on the unborn issues by its vaticination people can strategies their business and plan consequently. Grounded on the results a static model will be prepared. eg outliers models, Time series models, cast models.

b) Descriptive Method

The descriptive model classifies guests or prospects into groups grounded on the assaying relationship between the data as in Clustering, Summarization, Association rule, Sequence discovery, etc.

c) Summarization Technique

It is of good denotation in data mining. It's like the summary or determination the data. In data mining summarization can be befitted by utilizing various styles like medium, standard aside etc. This gives the general facts of the concentrated data.

d) Association Method

Association is an unsupervised machine learning approach. [11] It focuses on the association or catching on the collaborations between two. predicated on the symptoms and complaints a relationship gets formed to diagnose the complaint. In this it focuses on the frequent particulars which can take together like the people experiencing with covid will take the drug of fever, cough and nasal drops. Apriori, Eclat and F- P growth algorithm are the types of association algorithms.

e) Classification Algorithms

Association is an unsupervised machine learning approach. [8] It focuses on the association or catching on the collaborations between two. predicated on the symptoms and complaints a relationship gets formed to diagnose the complaint. In this it focuses on the frequent particulars which can take together like the people experiencing with covid will take the drug of fever, cough and nasal drops. Apriori, Eclat and F- P growth algorithm are the types of association algorithms.

f) Clustering Technique

Clustering is an unsupervised machine learning literacy grounded algorithm which consists of set of data points and form into clusters. analogous type of data objects belongs to the one group and the subset of this group is known as cluster. Clustering ways principally are of two types hard clustering and soft clustering. In hard clustering one data point can belong to one cluster only but in soft clustering it can belong to further than one cluster also.

g) Trend analysis

After investigating the data that what happed in the history it suggests that what will be in the future. currently as the number of cases are reducing and hence it follows the declining or down trend.

h) Regression Analysis

It's a statistical procedure for evaluating the couplings between the dependent variables or one or further independent variables. In this running values or pasture of numerical values can be forecast

i) Decision tree

The name shows the tree shaped structures of opinions. In this all the branches gives their results and the stylish one to draw out as the decision for the bracket of dataset. It includes ID3, C4.5 and wain (bracket and retrogression) Trees.

B. Healthcare methods used in data mining system

There are generally two types of supervised literacy and unsupervised literacy. The two approaches differ only in labeled or unlabeled datasets, since supervised literacy models use labeled input and event data, while unsupervised literacy models find themselves unlabeled datasets.

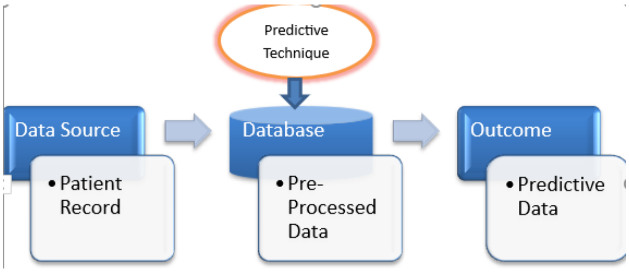

Figure 3: Healthcare Techniques

Brackets are a supervised literacy approach, while clustering is an unsupervised literacy approach. The results of this study show that decision trees are more fashionable predictors than others. We can understand by the Figure3.

V. USING TECHNIQUES IN HEALTHCARE

A. Artificial Neural Network (ANN)

This technique utilizes the brain as a base to develop algorithm for complex patterns and ratiocination problems as in our brain neuron processes the information. There are technical protrusions called axon which allows neuron to transmit electric and chemical signal to another cells. Neuron admit these signals via root like extensions called as dendrites. In the same way ANN has billions of processing units (input and affair units) which are connected by bumps. These input units admit colorful information and neural network tried to learn about the knowledge produced as an affair. [10] There's also a back propagation in which the network works backward from affair to input units to acclimate weight of its connection until the difference between the factual affair and asked outgrowth shows lower error as possible. It's also applied for perfecting the case's complaint operation [3]. It helps to elect the applicable system which can be used in health care assiduity for stylish results [4] artificial neural network fashion is most common fashion used in major complaint area like cancer etc. according to a check of artificial intelligence operations [5].

B. Bayesian Classifiers or Local Buyers

Bayesian Probabilistic Interpretation as Partial Beliefs and Bayesian Estimation Computing Based on Proposition Validity on i) previous estimate of probability ii) New applicable substantiation. The posterior estimation is done. It's grounded on Bayes theorem and bayes theorem find the probability of the individual thesis in the given data which can be calculated as,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Using Bayes theorem, we can find the probability of A passing, given that B has passed. Then, B is the substantiation and A is the thesis. The supposition made then' s that the predictors features are independent. That's presence of one particular point doesn't affect the other. Hence, it's called naive.

P (B) is the likelihood of the data.

[12] Naïve Bayes isn't a single algorithm but family of algorithms where every brace of features being classified as independent of each other and it's a family of simple "Probabilistic classifiers" and it's a supervised literacy algorithm. There are numerous naïve bayes classifiers like

Gaussian naïve bayes classifier, Multinomial naïve Bayes and Bernoulli naïve bayes.

C. Genetic algorithm

Optimization ways that use processes similar as inheritable combination, mutation, and natural selection in a design grounded on the generalities of natural elaboration.

D. Decision Tree

Tree shaped structures that represent sets of opinions. These opinions induce rules for the bracket of a dataset. Specific decision tree styles include Bracket and Retrogression Trees (wain) and Chi Square Automatic Interaction Detection (CHAID), wain are decision tree ways used for bracket of a dataset. In decision tree, there are three types of bumps chance bumps, decision bumps and end bumps represented by circle, square and triangle independently. The decision tree starts with single knot and also splits into its branches and also further splits into end bumps. To do overfitting there's pruning in which decision tree removes the corridor of tree which aren't of important use for decision timber. In this the pruning can be done by replace its branches with splints. There are two types of pruning: pre-pruning and post-pruning. This pruning is continued until the delicacy cannot be further improved. The colorful decision tree algorithm is listed below.

- Classification and regression tree (CART)
- Iterative Dichotomiser
- C4.5
- Chi-Square Automatic Interaction Detector. In short we can say CHAID.

E. CART

Bracket and retrogression tree predicts the value and at the end shows prophetic outgrowth. It takes the prophetic value grounded on other issues.

F. Iterative Dichotomiser

It's introduced by Ross Quinlan used to take decision from data set and it's the precursor of C4.5 which takes its affair as input for making opinions.

G. The C4.5 Algorithm

If we talking about decision tree algorithm, then C4.5 is one of the algorithms. It's also known as J48 in WEKA tool and as its decision tree can be used for the bracket and that's why it's known as statistical classifier. In this the regularized trait selection measure is known as gain rate and which can be measured as Gain(A) Split word(A) = Gain rate(A) where, A shows the trait in a data set D. It's relatively time effective and can be used for erecting more accurate decision trees.

H. Random Tree Classifier

It's a Machine Learning (ML) oriented algorithm and does ensemble bracket. In this it takes the decision from all the branches and the loftiest no. of vote vaticination will be the final decision tree. It's also time saving and effective classifier.

I. CHAID

Chi-squared automatic commerce sensor produces multiple branches of single or parent knot and it's constantly used for descriptive analytics

VI. REVIEW FROM RESEARCHES

Table 1: Equivalent study of given predictions in healthcare

S.N.	Techniques	Area
1	K- Nearest Neighbour	Diabetes, Cancer
2	SVM classifier	Diabetes
3	Apriori algorithm	Chronic Disease
4	Artificial Neural Network	Breast cancer
5	Logistic Regression	Breast cancer
6	Artificial Neural network	Chest Disease
7	Decision Tree	Diabetes
8	Naïve Bayes	Heart Disease
9	Neural Network	Eye disease
10	Decision tree induction method	Breast cancer
11	Decision tree (SPSS)	Diabetics
12	SVM	Stroke mortality in brain
13	Neural networks	Heart disease

Table 2: Some Techniques Area

S.N.	Techniques Area
1.	Nearest Neighbour Diabetes, Cancer
2.	SVM classifier Diabetes
3.	Apriori algorithm Chronic Disease
4.	Artificial Neural Network Breast cancer
5.	Logistic Regression Breast cancer
6.	Artificial Neural network Chest Disease
7.	Decision Tree
8.	Naïve Bayes
9.	Neural Network Eye disease
10.	Decision tree induction method
11.	Decision tree (SPSS) Diabetics
12.	SVM
13.	Neural networks Heart disease
14.	K- Nearest Neighbour Diabetes, Cancer

VII. CONCLUSION

In recent times there has been sharp hike in quantum of medical data in the field of medical care which is gathered by colorful electronic means and this along with rise in vacuity of provident and reliable computing outfit has encouraged numerous experimenters to start exploring these collected data. still, despite all available data our healthcare system is still worsening which shows that these data haven't been duly used. This paper explores the operation of a data mining method for effective use of medical data for complaint inoculation, which can make opinions at the right time and make people healthy. It has been observed that some data mining methods have been used on medical data before, while others may give better results in the future. The large amount of data available in the medical field makes it imperative to use data mining methods to make decisions and predictions in the medical field like identifying types of diseases. Vacancies in colorful medical facilities in different medical centers. On J48, a decision tree in Weka showing poor outcomes in the field of healthcare. This paper examines a variety of methods that differ from other methods, and it can be seen that arbitrary wood classifiers show effective results. These methods can be applied to a rich and varied subfield of healthcare, as well as to predict and better prepare for upcoming situations.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] Baek J W and Chung K 2020 , "Context deep neural network model for predicting depression risk using multiple regression ", IEEE Access 8 pp 18171-81

[2] Khan F A, Zeb K A M, Derhab A, Bukhari S A C 2021, "Detection and Prediction of Diabetes using Data Mining A Comprehensive Review IEEE Access"

[3] Delen D, Walker G, Kadam A 2005 "Predicting breast cancer survivability: a comparison of three data mining methods Artificial intelligence in medicine 34(2) pp 113-27"

[4] Luo L, Luo L, Zhang X, He X 2017, "Hospital daily outpatient visits forecasting using a combinatorial model based on ARIMA",models BMC health services research 17(1) pp 1-13

[5] B Taneja A 2013, " Heart disease prediction system using data mining techniques", Oriental Journal of Computer science and technology 6(4) pp 457-66

[6] B Krishnaiah V, Narsimha G, Chandra N S 2016, "Heart disease prediction system using data mining techniques and intelligent fuzzy approach a review ," International Journal of Computer Applications 136(2) pp 43-51

[7] Mahmoud H, Abbas E, Fathy I 2018, "Data mining and ontology-based techniques in healthcare management", International Journal of Intelligent Engineering Informatics 6(6) pp 509-26

[8] Naveenkumar S, Kirubhakaran R, Jeeva G, Shobana M, Sangeetha K Smart, "Health Prediction Using Machine Learning"

[9] Kumar H and Singh N 2017, "Review paper on Big Data in healthcare informatics", International Research Journal of Engineering and Technology 4(2) pp 197-201

[10] Er O, Yumusak N, Temurtas F 2010, " Chest diseases diagnosis using artificial neural networks", Expert Systems with Applications 37(12) pp 7648-55

[11] Tang P H and Tseng M H 2009, " Medical data mining using BGA and RGA for weighting of features in fuzzy k-NN classification", In 2009 International Conference on Machine Learning and Cybernetics IEEE 5 pp 3070-75

[12] Balakrishnan S and Narayanaswamy R 2009 Feature selection using fcbf in type ii diabetes databases International Journal of the Computer the Internet and the Management 17(1) pp 50-8

[13] Chaurasia V and Pal S 2013, " Early prediction of heart diseases using data mining techniques", Caribbean Journal of Science and Technology pp 208-17

[14] Bahrami B and Shirvani M H 2015 "Prediction and diagnosis of heart disease by data mining techniques", Journal of Multidisciplinary Engineering Science and Technology (JMEST) 2(2) pp 164-8

[15] Anu Sharma, M.K Sharma, Rakesh Kr. Dwivedi, "Exploratory data analysis and deception detection in news articles on social media using machine learning classifiers", Ain Shams Engineering Journal, Volume 14, Issue 10, 2023, 102166, ISSN 2090-4479.

[16] Anu Sharma et.al, "Literature Review and Challenges of Data Mining Techniques for Social Network Analysis," Advances in Computational Sciences and Technology, 2017.