How to Cite:
Burnet, D. (2019). Numbers of approaches of data mining are being used by various researchers in healthcare sector. *Tennessee Research International of Social Sciences, 1*(1), 1–11. Retrieved from https://triss.org/index.php/journal/article/view/3

Numbers of approaches of data mining are being used by various researchers in healthcare sector

Dustine Burnet
Vanderbilt University, Nashville, United States

***Abstract*---**This research will help summarize all the techniques used in health care along with their accuracy level. In this, we understand the application of data mining in the health care sector to extract useful information to predict disease using data mining applications is a difficult task. Several approaches to data mining are being used by various researchers in the healthcare sector, so the number of techniques in this field is increasing. It helps to significantly reduce the human effort and increase diagnosis accuracy can develop the style of data mining to reduce cost and time constraints in terms of resources and human expertise. This paper has been presented to demonstrate how a certain application challenge can be addressed with the help of the valuable tool of data mining.

***Keywords*---**Costumer, Healthcare sector, Researchers, Disease, Human effort.

Introduction

Data mining is all about discovering useful information from a large data repository. It involves data selection, exploration, and building models to find unknown patterns. It deals with the analysis of data and then processes it into useful information. Actually, data mining deals with the finding of various Correlations or patterns in the large relational database Systems of the Hospital Management Software. These Techniques help the administrators in the analysis and also creates a scope for the discovery of various unsuspected relationships among their database. This in turn is very helpful in making useful decisions. They have been used extensively by many organizations. Data mining is gaining popularity in recent times and is also becoming increasingly essential for several other fields.

Tennessee research international of social sciences © 2019.
ISSN: 2766-7464 (Online)
Publisher: Smoky Mountain Publishing
Manuscript submitted: 09 March 2019, Manuscript revised: 18 April 2019, Accepted for publication: 27 May 2019

Due to modernization in record keeping, a huge amount of raw data is generated by healthcare organizations daily. These data are very complex and huge. Thus it is difficult to analyze them with the help of the traditional tools. There is a lack of effective tools that can help trigger the detection of new trends in the healthcare systems. In the Existing System, the data is being analyzed manually. Due to this many relationships that are hidden but prove to be potentially useful may not be recognized by the Hospital Administrators. Most of the money spent by the Government health Organisations go waste due to improper treatments and anomaly in the detection of the exact cause of the disease. People cannot use this huge amount of data effectively for the benefit of the patients. Most of the money spent by the Government health Organisation is wasted. This happens due to diagnostic Errors such as over-testing, Hospital Re-Admissions, and Medical Errors, claim Processing.

Proposed System

Data mining technology enables computers to assist hospital administrators to find valuable information and data from massive data and assist management personnel in making decisions. The main object used for the discovery of association rules is a transactional database. Its purpose is to make the distance between individuals belonging to the same category become as short as possible & vice versa. In the hospital management system, data mining techniques can be used for finding the hidden patterns which help make more effective decisions concerning the health of the patients. This paper will help understand, review, and analyze different data mining techniques such as clustering, classification, and Regression, etc used in the hospital Management Systems or the Healthcare units. Few Algorithms will be used to classify data effectively which are as follows:

1) K-Nearest Neighbors (KNN)
2) Decision Tree (DT)
3) Artificial Neural Network
4) Support Vector Machine (SVM)

Advantages of Proposed System

This avoids searching for the required data from the database which helps save time and resources, to obtain long-term, systematic and comprehensive data. This system provides great support to the scientific management of the hospital. It promotes the extraction of knowledge that can enable support for costsaving and an effective decision making procedure to help serve the patients better. This system helps Healthcare insurers detect fraud and abuse. By Adopting this system, Healthcare organizations will be able to make effective customer relationship management decisions. It becomes easy for Physicians to identify effective treatments and best practices for the Patient. In this way, Patients receive better and more affordable healthcare services.

Data Mining

Data mining is being extensively used in various fields due to its limitless approaches to mine the data in an extremely target-oriented manner. Initially, the storage of data followed by the extraction of useful information was a very difficult

task. But later on, with the advancement and research in the field of technology, storage and extraction of data became a very easy task. Data mining consist of 5 major elements, first is to extract, transform, and load data onto the data warehouse system. The second is to store and manage the data in a multidimens0ional database. The third is to provide data access to analysts. Fourth is to analyze the data by application software. The fifth is to finally present the data in a useful format.

Steps in a Data Mining Process are:

1) Data is collected and integrated from various sources.
2) The selection of data is done based on the given criteria.
3) The Data collected may be inconsistent as it might have errors. All the errors need to be removed.
4) The data obtained even after pre-processing may not be ready for the mining procedure and thus there is a need for it to be transformed into a form that is appropriate for the mining procedure to be carried out.
5) Meaningful patterns are extracted from large data. At last, meaningful patterns help in decision making.

Classification

Classification maps data into predefined groups or classes where the classes are determined before examining the data. They often describe these classes by looking at the characteristics of data attribute values already known to belong to the classes. The main aim of this technique is to accurately predict the target class for each data value. To finalize as to which category the particular data value belongs; it needs to process training and a predictive set. It first develops relationships between the attributes of the training data set. Then it has to be given a predictive data set, which will possess similar attributes but with different data values. Then it analyzes the given data in the training set to compute the values in the predictive set by placing the different data values in different classes based upon the relationship of data attributes.

Table1: Training Set

AGE	HEART RATE	BLOOD PRESSURE	HEART PROBLEM
62	79	145/70	Yes
35	82	115/75	Yes
79	65	110/68	No

Table 2: Predictive Set

AGE	HEART RATE	BLOOD PRESSURE	HEART PROBLEM
45	96	143/69	?
63	54	108/73	?
83	95	115/68	?

This technique makes use of some predictive rules that are usually expressed in the form of IF-THEN rules where the IF part consist of the conjunction of conditions and the THEN part predicts a certain prediction attribute value that satisfies the first part. The rule predicting the first row in the training set may be represented as follows:
IF (age>60 and blood pressure>140/70) or (age=62 and heart rate>72) then Heart problem=yes.
The above-mentioned technique has been experimented and has been concluded to provide around 80% accurate prediction rates.

Decision Tree

A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value and each branch represents an outcome of the test. The tree leaves represent classes or in other words, the various class distributions involved. They are used to predict the membership of objects to different categories (classes), taking into account the values that correspond to their attributes. The decision tree has predicted correct results for 86.25% of cases. So we can use decision trees to obtain this useful information to make important decisions in the Hospital Management System.

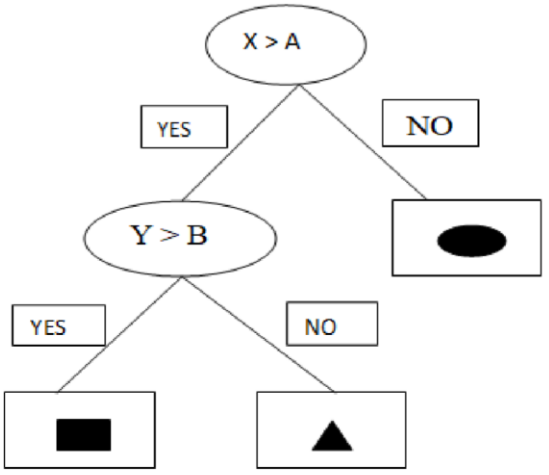

Fig 1: Structure of a Decision Tree

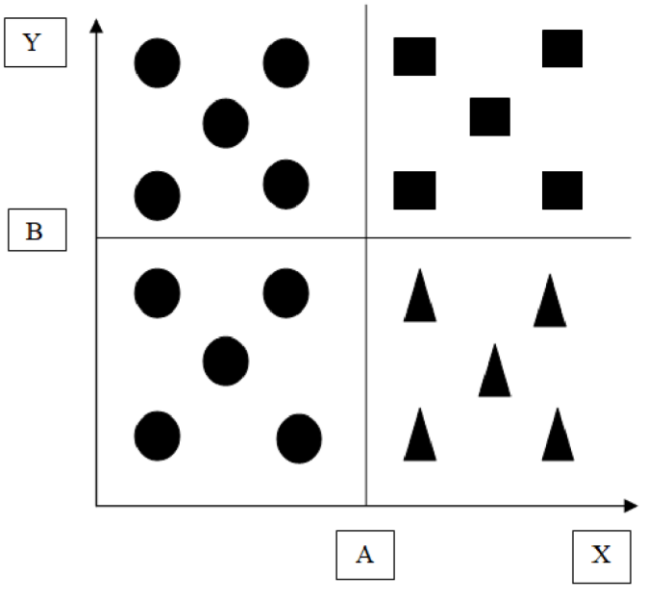

Fig 2: Graphical Representation of a Decision Tree

Support Vector Machines

The Support Vector Machines work as the linear separator between two entirely different data points to identify two different classes in the multi-dimensional environment. SVM uses a very big set of non-linear features that is task-independent. The main idea behind this approach is to maximize the margin between the classes and to minimize the distance between the points in the hyperplane. SVM splits the dataset into two vector sets under 'n' dimensional

space vector. It can be easily implemented with the help of different kernel functions. Its accuracy is better than all other available techniques as it has 91%. The kernel functions are:

a) Polynomial kernel with degree d $K(X,Y) = (XT Y + 1)d$
b) Radial basis function kernel with width σ $K(X,Y) = \exp(- \|X - Y\|2 / (2 \sigma 2))$

   It is very closely associated to radial basis function in the neural networks. The feature space is infinite-dimensional.
c) Sigmoid with parameter θ and. $K(X,Y) = \tanh(kxTy + \theta)$ The'd', 'σ' and 'θ' are parameters chosen by the user. The Algorithm is as follows:
   1. Step 1: Prepare the pattern matrix
   2. Step 2: Select the kernel function to use.
   3. Step 3: Select the parameter of the kernel function and the value of C [use the values suggested by the SVM software]
   4. Step 4: Execute the training algorithm and obtain the αi for further computation.
   5. Step 5: Unseen data can be classified using the αi and the support vectors.

Neural Network Scheme

In the past, the neural network was considered as the best classification algorithm before the introduction of the decision tree and Support Vector Machines. It has been used as the algorithm supporting the diagnosis of diseases like cancer. In this, basic elements are nodes or neurons. These neurons are connected to form a Network like Structure. Within this network, they work together to produce the output functions. An activation number is associated with each neuron and a weight is assigned to each edge within the Neural Network. The basic property of these Networks is that it can minimize the error by adjusting its weights and by making changes in its structure as it is adaptive in its nature.

One big advantage of this technique is that it can handle noisy data effectively for training and can prove useful in classifying a new type that is entirely different from the training data. They are considered to be tolerant of any sort of fault as they can easily build new observations from the existing ones. These new observations generated prove to be very useful in those situations where some neurons within the network happen to fail. There are several disadvantages to Neural Networks. First, it needs many parameters like the optimum number of hidden layer nodes. Its performance to classify data effectively is very sensitive to the parameters that have been selected. Moreover, the training or the learning process associated with the Neural Network is very slow and also expensive.

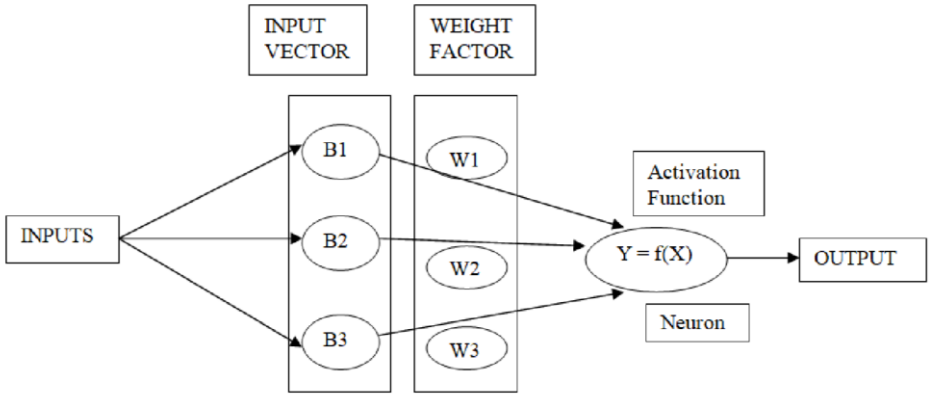

Fig 3: Structure of a Single Neuron

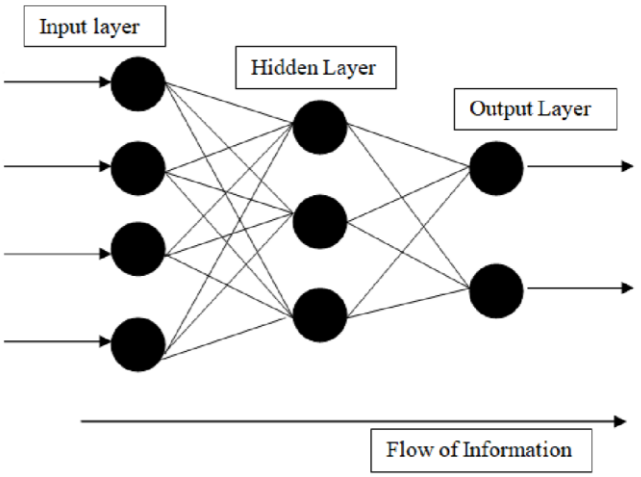

Fig 4: Neural Network Scheme

K-Nearest Neighbors

K-Nearest Neighbors (KNN) is one of the simplest technique which deals with the discovery of the identity of an unknown point by using the knowledge of the data points that are present in its nearest neighbors. It has several applications in various fields like the analyses of health database of patients, the study of fields of images, and pattern recognition. All of these make use of the K-Nearest Neighbors technique to analyze distinct information for LDA which helps in the classification of chronic diseases to generate an early warning system It is extensively used to analyze the relationship between the heart and the artery of the patients, high blood pressure levels and various factors associated with the risk for chronic diseases to build a system based on an early warning to reduce the incidence of complications of these diseases.

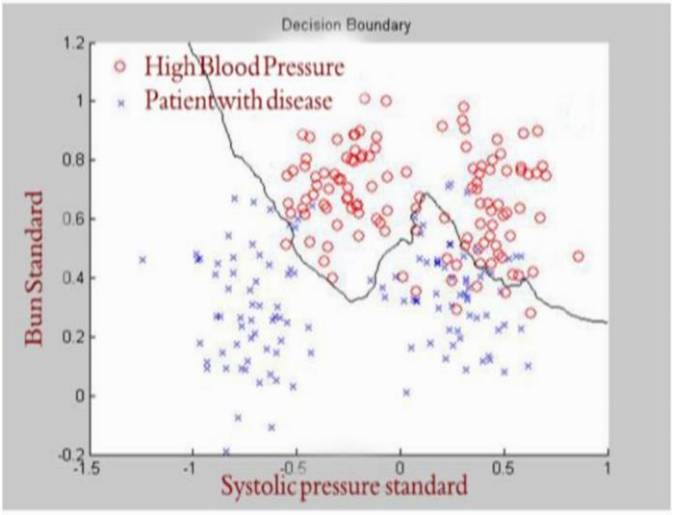

Fig 5: Application of K-Nearest Neighbor Algorithm

Clustering

It is usually determined as a much-unsupervised learning technique that is different from the classification technique (supervised learning method). It is best suited for large amounts of data. It works by observing independent variables. The main task is to form clusters from large databases based on the similarity measure. Different types of clustering algorithms are defined and the various clustering algorithms used in health care are described below.

Table 3: Clustering Mechanisms & Description

Technique	Description
1. Partitioned Clustering	With the help of 'n' data points maximum possible of 'k' clusters is obtained by relocating objects to 'k' clusters.
2. Hierarchical Clustering	Data points are partitioned in tree form either top-down or bottom-up.
3. Density-based Clustering	It can handle cluster of any arbitrary shape whereas above two can handle only spherical shape clusters.

Association Rules

Association in data mining refers to the task of uncovering relationships among data. This can be done by building up a few association rules. An association rule is a model that helps to identify the specific type of associations within the data. One of the main aims of these methods is to help the administrators in the hospital understand the description of the various patterns held within the data very clearly. This can be achieved by making use of the association rules. These

Association rules in relational databases relate the presence of values of some attributes with values of some other attributes in the same tuple. The rule [X = x])] [Y = y] says that whenever the attribute X takes value x in a tuple, the attribute Y takes value y in the same tuple. The Discovery of these rules is one of the main techniques that can be used by both physicians as well as managers to obtain knowledge from large and extensive medical database systems. For Example, Doctors will find it more appropriate to describe his knowledge utilizing rules like “if fever is high and cough is moderate then the disease is A” rather than by making use of the rules such as “if fever is 38.799C and cough is 6 over 10 then the disease is A”. Thus in recent times, some people apply semantics to improve the mining-related to these association rules from a database that contains more precise and accurate values.

Regression

Regression is a special technique in the field of data mining that helps to identify those functions that are useful to demonstrate the correlation among different variables. It is a tool and can be easily constructed using training data sets. It can be further classified into linear and nonlinear based upon the value of a certain count of independent variables. To derive an estimate of the association between two types of variables in which one is dependent and the other is an independent one, linear regression will be used. One of the disadvantages of this technique is that it cannot be used for categorized data.

Table 4: Different Regression Techniques

Technique used	Purpose
Weighted SV Regression	To provide better healthcare services by continuously monitoring patients.
Regression decision tree algorithm	To study number of hospitalization days.
Linear regression	For effective utilization of hospital resources.

Accuracy Analysis of These Techniques

There are several challenges in the processing of the data that comes from the hospital or any healthcare System as they can create several serious obstacles in the decision-making process. Different hospital systems use different formats for storing and managing the data. Many of them even Lack the standard forms of data that can be used for storage. Mostly, both medical organizations like the Hospitals & Other Healthcare units and the patients are not ready to share their private information. Another reason is that building a centralized data warehouse is extremely time-consuming and at the same time an expensive process.

Table 5: Accuracy Analysis of Each Technique

Disease	Data Mining Technique	Algorithm	Accuracy level (%)
Heart Disease	Classification	Naïve	60
Cancer	Classification	Decision Table	97.77
HIV AIDS	Classification, Association Rule mining	J48	81.8
Blood Bank	Classification	J48	89.9
Brain Cancer	Clustering		85
Tuberculosis	Naïve Bayes Classifier	NN	78
Diabetes	Classification	C4.5	82.6
Kidney Dialysis	Classification	Decision making	75.97
Dengue	Classification	C5.0	80
Hepatitus C	Classification	Decision tree	73.2

Conclusion

Several approaches to data mining are being used by various researchers in the healthcare sector, so the number of techniques in this field is increasing. This research will help summarize all the techniques used in health care along with their accuracy level. In this, we understand the application of data mining in the health care sector to extract useful information to predict disease using data mining applications is a difficult task. It helps to significantly reduce the human effort and increase diagnosis accuracy can develop the style of data mining to reduce cost and time constraints in terms of resources and human expertise. This paper has been presented to demonstrate how a certain application challenge can be addressed with the help of the valuable tool of data mining.

References

Ahmad, P., Qamar, S., & Rizvi, S. Q. A. (2015). Techniques of data mining in healthcare: a review. *International Journal of Computer Applications*, *120*(15).

Bellazzi, R., & Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. *International journal of m* Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119.edical informatics*, *77*(2), 81-97.

Bellazzi, R., Diomidous, M., Sarkar, I. N., Takabayashi, K., Ziegler, A., & McCray, A. T. (2011). Data analysis and data mining: current issues in biomedical informatics. *Methods of information in medicine*, *50*(6), 536.

Chen, H., Fuller, S. S., Friedman, C., & Hersh, W. (Eds.). (2006). *Medical informatics: knowledge management and data mining in biomedicine* (Vol. 8). Springer Science & Business Media.

Cheng, L., Liu, F., & Yao, D. (2017). Enterprise data breach: causes, challenges, prevention, and future directions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(5), e1211.

Cios, K. J., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial intelligence in medicine*, *26*(1-2), 1-24.

Cios, K. J., Pedrycz, W., & Swiniarski, R. W. (1998). Data mining and knowledge discovery. In *Data mining methods for knowledge discovery* (pp. 1-26). Springer, Boston, MA.

Duan, L., Street, W. N., & Xu, E. (2011). Healthcare information systems: data mining methods in the creation of a clinical recommender system. *Enterprise Information Systems*, *5*(2), 169-181.

Durairaj, M., & Ranjani, V. (2013). Data mining applications in healthcare sector: a study. *International journal of scientific & technology research*, *2*(10), 29-35.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37-37.

Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. *Journal of healthcare information management*, *19*(2), 65.

Kumar, S., & Singh, M. (2018). Big data analytics for healthcare industry: impact, applications, and tools. *Big Data Mining and Analytics*, *2*(1), 48-57.

Leung, M. W., Yen, I. H., & Minkler, M. (2004). Community based participatory research: a promising approach for increasing epidemiology's relevance in the 21st century. *International journal of epidemiology*, *33*(3), 499-506.

Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I. A. T., Siddiqa, A., & Yaqoob, I. (2017). Big IoT data analytics: architecture, opportunities, and open research challenges. *IEEE Access*, *5*, 5247-5261.

McGhin, T., Choo, K. K. R., Liu, C. Z., & He, D. (2019). Blockchain in healthcare applications: Research challenges and opportunities. *Journal of Network and Computer Applications*, *135*, 62-75.

Siemens, G., & Baker, R. S. D. (2012, April). Learning analytics and educational data mining: towards communication and collaboration. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 252-254).

Soni, J., Ansari, U., Sharma, D., & Soni, S. (2011). Predictive data mining for medical diagnosis: An overview of heart disease prediction. *International Journal of Computer Applications*, *17*(8), 4348.

Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, *2*(02), 250-255.

Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. *International Journal of Bio-Science and Bio-Technology*, *5*(5), 241-266.

Yang, Q., & Wu, X. (2006). 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, *5*(04), 597-604.

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems*, *36*(4), 2431-2448.