Data Mining In Healthcare Industry: A Systematic Research on Recent Studies

Divyanshu Keshari[1], Prof. Durgesh Pandey[2]
[1] Dept of Computer Science and Engineering
[2] Guide, Assistant Professor, Dept of Computer Science and Engineering
[1, 2] Saraswati higher education and technical College of engineering, babatpur Varanasi Uttar Pradesh.

Abstract- *The introduction of data-use technologies and methodologies has enhanced knowledge discovery in databases. One of the most important stages in data mining. The process of detecting and extracting patterns from huge volumes of data is known as data mining. Both the data mining and healthcare industry have produced some of the most accurate early detection systems and other healthcare-related solutions using clinical and diagnostic data. We analysed the various works in this field in terms of approach, algorithms, and conclusions in respect to this rising concern. The works examined in this review study have been grouped according to disciplines, models, tasks, and approaches. To complete the study, the outcomes and evaluation methodologies of selected research are analysed, and a summary of the findings is provided.*

Keywords- Data Mining Data Mining in Healthcare Health Informactics

I. INTRODUCTION

It is incredibly expensive to keep data or information. However, thanks to advancements in information collecting methods and the WWW during the previous twenty-five years, a vast quantity of information or data is now accessible in electronic format. Database sizes are continually increasing in order to hold such a big quantity of data or information. Such databases include a wealth of important information. This knowledge might be extremely beneficial in any decision-making process. It is made feasible by the use of data mining or knowledge discovery in databases. Data mining is the process of collecting usable information from a big collection of previously unknown data.

In the healthcare industry, five stages of data mining have been identified

- **Selection:** The data is picked based on a set of criteria in this stage. For example, if all of those people own a bicycle, we may create data subsets in this way.
- **Preprocessing:** This step removes extraneous information, such as recording a patient's gender when administering a pregnancy test. It is also known as the data cleaning phase.
- **Transformation**: This stage only changed the data that is useful in a certain study, such as data related to a specific demography in market research.
- **Data mining:** Data mining is a phase in the discovery of knowledge. This step is critical for detecting meaningful data trends.
- **Interpretation and evaluation:** The system's observed relevant patterns are interpreted into knowledge at this level. This knowledge may then be used to make good decisions.

The Importance and Uses of Data Mining in Medicine and Public Health

Despite methodological differences and disagreements, the health business currently has a growing need for data mining. There are several reasons for promoting the use of data mining in the health industry, encompassing both public and private health issues (which, in fact, as can be shown later, are also stakeholders in public health). Information overload. Computerized health records may provide a wealth of information. The huge volume of data stored in these databases, on the other hand, makes it very difficult, if not impossible, for humans to sift through and unearth information. Indeed, some experts believe that medical progress has stalled, blaming the prohibitively large and complicated amount of medical information available today. This work is best suited to computers and data mining. (Evidence-based medicine and avoiding hospital errors.) When medical organisations use data mining on their existing data, they may discover new, useful, and even life-saving information that would otherwise lie dormant in their databases. For example, a current hospital safety study revealed that around 87 percent of hospital deaths in the United States might have been averted if hospital staff (including doctors) had been more careful in preventing errors (Health Grades Hospitals Study 2007). By mining hospital data, hospital management and government authorities may be

able to identify and correct such safety issues. Policy development in public health. Larva et al. (2007) used GIS and data mining technologies such as Weka and J48 to examine similarities across community health centres in Slovenia (free, open source, Java-based data mining tools). Using data mining, they were able to identify similarities among health institutions, which led to policy recommendations to their Institute of Public Health. They concluded that "data mining and decision support technologies, including novel visualisation methodologies, may lead to enhanced decision-making performance."

Health care industry

Healthcare companies provide clinical services, manufacture pharmaceuticals and medical equipment, and provide healthcare-related support services such as medical insurance. It is sometimes referred to as the medical industry. These companies play an important role in the diagnosis, treatment, nursing, and management of illness, disease, and accidents.

The healthcare industry also provides patients with preventive, remedial, and therapeutic services. To offer these services, doctors, nurses, medical administrators, government agencies, pharmaceutical firms, medical equipment manufacturers, and medical insurance companies must collaborate.

Data mining challenges in healthcare

As we all know, a vast quantity of healthcare data is generated and maintained by different healthcare facilities. However, there are a number of difficulties with healthcare data that might make it difficult to make effective decisions. The first issue with healthcare data is that the format in which it is stored differs across healthcare facilities. So yet, no common data storage format has been defined. In epidemic situations, the lack of a uniform framework may aggravate the issue. Assume that a disease outbreak develops over a country's numerous geographical regions.

Objective of the study

1. To list current applications and emphasize the significance of data mining in medical and public policy.
2. Health, in order to identify data mining methods utilized in other disciplines that may be implemented in the health sector.
3. Identifying concerns and obstacles in data mining as they apply to medical practice.
4. Outline some guidelines for using data mining to uncover information in electronic databases.

Scope of the study

Increasing computer-based data analysis understanding, online educational availability, and building an integrated learning method among medical practitioners would undoubtedly aid in correct diagnosis and an efficient treatment management plan in India. Innovative medical technologies are critical for patient care. This is also true for the prevention of numerous illnesses connected to cleanliness, communicable diseases, and addiction-associated disorders such as lung cancer, mouth cancer, liver cirrhosis, and so on. In the future, the extent of technological applications such as data mining techniques-based systems in India's healthcare system would really create tremendous changes at every level. Today, the internet serves as a portal to global information as well as a vast platform for national media and documentation, which will be very beneficial in the future deployment of data mining methods.

II. LITERATURE REVIEW

2.1 Joti Sony et.al "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction" (2011)

The successful use of data mining in high-profile disciplines such as e-commerce, marketing, and retail has led to its use in other businesses and sectors. Healthcare is one of these areas that is still in its infancy. The healthcare environment continues to be "information rich" yet "knowledge poor." Within healthcare systems, there is a lot of data. However, there is a scarcity of good data analysis tools for uncovering underlying correlations and patterns. This study work aims to present an overview of current strategies for knowledge discovery in databases employing data mining techniques that are used in today's medical research, specifically in Heart Disease Prediction. A number of experiments have been carried out to compare the performance of predictive data mining techniques on the same dataset, and the results show that Decision Tree outperforms, and sometimes Bayesian classification has similar accuracy as Decision Tree, but other predictive methods such as KNN, Neural Networks, and Classification based on clustering do not perform well. The second result is that the accuracy of the Decision Tree and Bayesian Classification increases even more after using a genetic algorithm to minimise the real data quantity in order to get the ideal subset of attributes required for heart disease prediction.

2.2 Umar Shaniqua et.al "Data Mining in Healthcare for Heart Diseases" (2015)

Data Mining is a field of study that involves the extraction of relevant information or knowledge from prior data. For data mining, several approaches are used. Data mining might be used in a variety of industries, including healthcare. Heart or cardiovascular illnesses are a major concern in the worldwide healthcare business. Many patients perished as a result of a lack of information. Because the healthcare business generates a large quantity of data, we may utilise data mining to uncover hidden patterns and fascinating information that can aid in effective and efficient decision making. Data mining in healthcare is a critical and complex process that must be completed correctly. It aims to tackle real-world health concerns in illness detection and treatment. This endeavour is also an attempt to discover intriguing patterns in cardiac patient data. Three algorithms are utilised in two separate contexts. Decision Tree, Neural Network, and Nave Bayes are the algorithms that have been implemented.

2.3 NishaJoti et.al "The Third Information Systems International Conference Data Mining in Healthcare – A Review" (2015)

Knowledge discovery in databases (KDD) is concerned with the creation of methodologies and procedures for using data. Data mining is a critical phase in the KDD process. Data mining is the process of discovering and extracting patterns from massive amounts of data. From clinical and diagnostic data, both the data mining and healthcare industries have developed some of the most accurate early detection systems and other healthcare-related tools. In relation to this emerging issue, we examined the many papers in this subject in terms of technique, algorithms, and findings. The works evaluated in this review paper have been consolidated in accordance with the disciplines, model, tasks, and methodologies. For chosen studies, the results and assessment techniques are examined, and a summary of the findings is offered to finish the study.

2.4 Diva Tamar et.al "A survey on Data Mining approaches for Healthcare" (2013)

Data mining is one of the most enthralling areas of study that is gaining traction in health care organisations. Data mining is vital for identifying new trends in healthcare organisations, which is beneficial to all parties involved in this industry. This review investigates the value of several Data Mining methods in the health area, such as classification, clustering, association, and regression. In this work, we provide a quick overview of various strategies, as well as their benefits and drawbacks. This report also emphasises Data Mining uses, problems, and potential concerns in healthcare. This study also discusses recommendations for selecting the best accessible Data Mining approach.

2.5 Anand Sharma et.al "Emerging Applications of Data Mining for Healthcare Management - A Critical Review" (2014)

In this study, we give a critical evaluation of the current research in data mining applications for healthcare management. The goal of this research is to look at new and developing areas of data mining methods utilised in healthcare management. Infection control surveillance, disease diagnosis and treatment, healthcare resource management, customer relationship management, fraud and anomaly detection, healthcare administration, hospital management, and public health are among the applications covered in this article. This study examines the data mining objectives attained, functions performed, and techniques employed in various applications.

.6 Srinivas et.al "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks" (2008)

The healthcare environment is often regarded as 'information rich' yet 'knowledge poor.' Within healthcare systems, there is a lot of data. However, there is a scarcity of good data analysis tools for uncovering underlying correlations and patterns. Knowledge discovery and data mining have a wide range of applications in business and science. The implementation of data mining methods in the healthcare system may provide valuable insights. In this paper, we look at how classification-based data mining approaches including rule-based data mining, decision trees, Nave Bayes, and artificial neural networks may be used to large amounts of healthcare data. The healthcare business gathers massive volumes of data, which are regrettably not "mined" to uncover hidden information. One Dependency Augmented Nave Bayes classifier (ODANB) and naïve creedal classifier 2 (NCC2) are utilised for data preparation and effective decision making. This is an extension of naive Bayes to imprecise probabilities that tries to provide robust classifications even with tiny or partial data sets. The discovery of hidden patterns and correlations is often underutilised. It can forecast the chance of individuals developing heart disease based on medical characteristics such as age, gender, blood pressure, and blood sugar. It allows for the establishment of substantial information, such as patterns and correlations between medical aspects associated to heart disease.

2.7 Muhammad Aurangzeb Ahmad, (2018) "Interpretable Machine Learning in Healthcare"

Increased requests for machine learning and AI-based systems to be controlled and held responsible in healthcare are accompanying the push for more machine learning penetration in healthcare. Machine learning models that are interpretable may help to keep machine learning systems responsible. Healthcare presents unique problems for machine learning since the criteria for explainability, model integrity, and overall performance are substantially greater than in most other areas. In this article, we examine the concept of interpretability in the context of healthcare, the different subtleties connected with it, interpretability difficulties that are particular to healthcare, as well as the future of interpretability in healthcare.

WHILE the application of machine learning and artificial intelligence in medicine dates back to the early days of the profession [1,] it is only in recent years that there has been a drive to recognise the need for machine learning-powered healthcare solutions. As a result, academics believe it is only a matter of time until machine learning becomes commonplace in healthcare.

2.8 K. Shailaja, (2018) "Machine Learning in Healthcare: A Review"

Machine Learning is a new and highly complex technical application that has become a big business trend. Machine Learning is everywhere and is extensively employed in a variety of applications. It is crucial in many industries, including finance, medical research, and security. Machine learning is utilised to detect patterns in medical data sources and provides great illness prediction skills. We discuss several machine learning methods utilised for generating effective decision support for healthcare applications in this research. This work contributes to closing the research gap in the development of effective decision support systems for medical applications.

Machine learning is a vast multidisciplinary field having roots in statistics, mathematics, data processing, and knowledge analytics, among other things, making it difficult to come up with a unique description [1]. ML is a kind of artificial intelligence that gathers knowledge from training data. We are not telling the computers where to look in this learning, and it is at the root of the tree with many branches and sub-branches.

2.9 Ankita R. Nambiar,(2017) "A Study of Machine Learning in Healthcare"

There have been major advancements in how machine learning may be employed in many businesses and studies during the last several years. This paper addresses the possibilities of applying machine learning technology in healthcare and describes numerous industry activities that are leveraging machine learning projects in the healthcare sector. Healthcare is one of the fastest growing industries today, and it is undergoing a thorough worldwide makeover and revolution. According to Russell Reynolds and Associates, global healthcare expenses, which are now estimated at $6 trillion to $7 trillion, are expected to exceed $12 trillion in only seven years. This tendency is also seen on a domestic level in the United States. Total healthcare expenditure in the United States climbed by 5.3 percent and has now surpassed $3 trillion nationally. Furthermore, healthcare expenditure in the United States accounts for 17% of total GDP; our healthcare expenses are growing at almost double the pace of our economic growth. Aside from an increase in consumer healthcare expenditure, the federal government has been obliged to pay more and more for healthcare as expenses grow too expensive for people to afford. The federal government's allocation for healthcare expenditure climbed by 11.7 percent in 2014, reaching an all-time high of $844 billion in 2015. This increase in government financing reflects the large discrepancy between the expense of healthcare and the financial burden on consumers.

2.10 B. Nithya, (2017)" Predictive Analytics in Health Care Using Machine Learning Tools and Techniques"

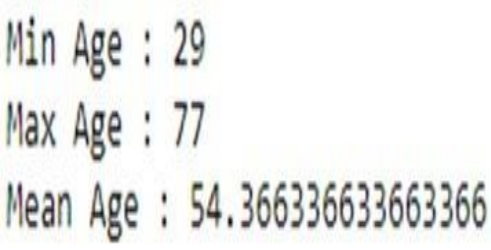
Machine learning is the way to go when we have a large data collection on which we want to do predictive analysis or pattern identification. Machine Learning (ML) is the most rapidly growing area in computer science, and health informatics is a major concern. The goal of Machine Learning is to create algorithms that can learn and improve over time and can be used to make predictions. Machine Learning procedures are extensively applied in a variety of areas, with the health care industry benefiting the most from machine learning prediction approaches. It provides a range of alerting and risk management decision support tools with the goal of increasing patient safety and healthcare quality. With the desire to lower healthcare costs and the shift toward customized treatment, the healthcare sector is confronted with issues in critical areas such as electronic record management, data integration, and computer assisted diagnosis and illness prediction. To solve these difficulties, machine learning provides a broad variety of tools, methodologies, and systems. This article examines numerous prediction approaches and tools for Machine Learning in practise. A look at the applications of Machine Learning in many sectors is also covered here, with a focus on its importance in the health care business.

2.11 MIN CHEN1,(2017) "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities"

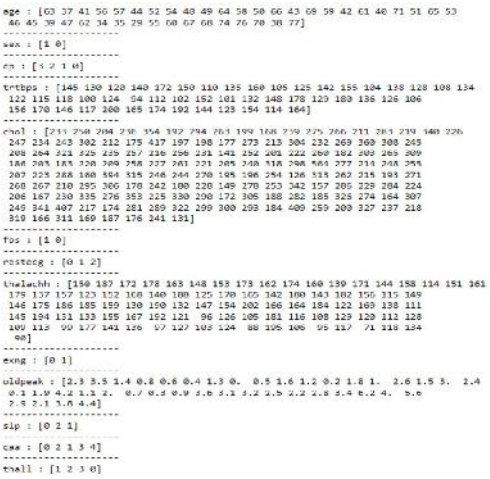
With the rise of big data in the biomedical and healthcare sectors, precise medical data analysis promotes early illness identification, patient treatment, and community services. When the quality of medical data is poor, the analytical accuracy suffers. Furthermore, various locations display distinct features of particular localized illnesses, which may make disease outbreak prediction difficult. In this study, we simplify machine learning methods for effective chronic illness outbreak prediction in disease-prone areas. We tested the updated prediction models using real-world hospital data obtained in central China between 2013 and 2015. We employ a latent component model to rebuild missing data to address the challenge of incomplete data. We conduct an experiment on a persistent localised illness of cerebral infarction. Using structured and unstructured hospital data, we present a novel convolutional neural network (CNN)-based multimodal illness risk prediction method. To the best of our knowledge, no current study in the field of medical big data analytics has focused on both data types. When compared to various conventional prediction algorithms, our suggested approach has a prediction accuracy of 94.8 percent and a convergence time that is quicker than the CNN-based unimodal disease risk prediction algorithm.

III. RESULTS & DISCUSSION

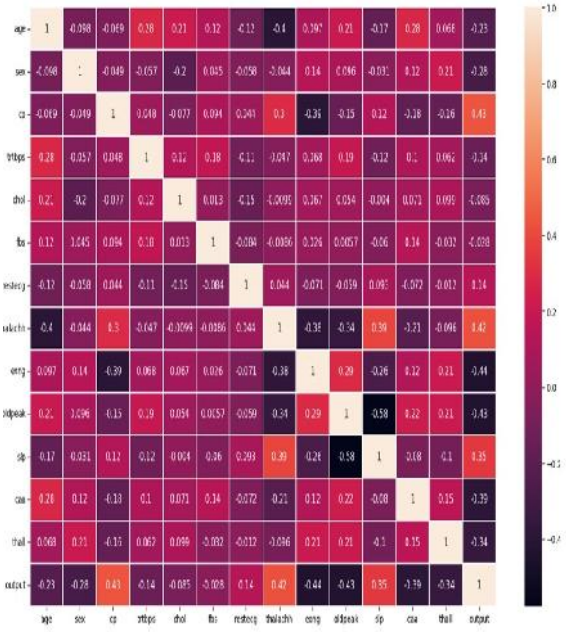
This chapter, a comparative study of data mining applications in healthcare sector by different researchers given in detail. Mainly data mining tools are used to predict the successful results from the data recorded on healthcare problems. Different data mining tools are used to predict the accuracy level in different healthcare problems.



At the time of Implementation, it is found that the healthcare issues regarding heart are mainly seen in adults who are between 29 to 77 age group. And if we calculate the mean age, we get 54.3663 values. So hence we can say that at the age 54 people starts facing health issues.

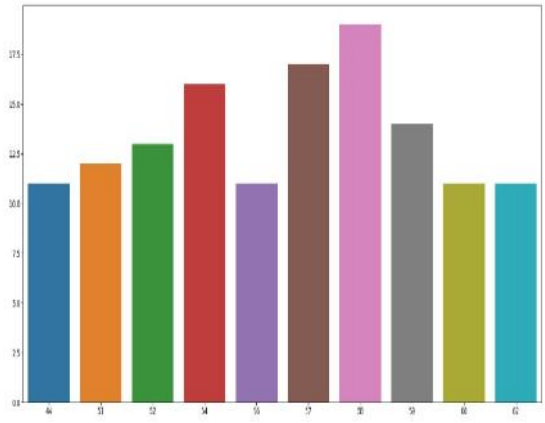


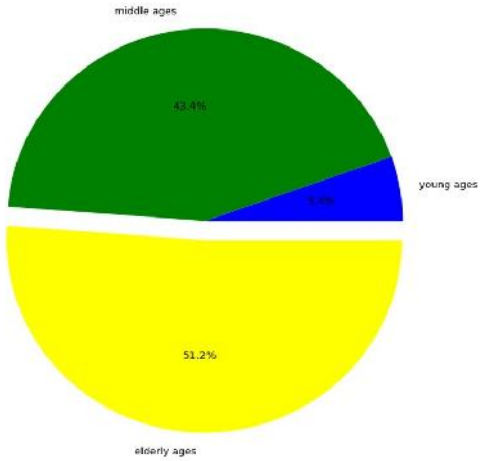
	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
5	57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
6	56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
7	44	1	1	120	263	0	1	173	0	0.0	2	0	3	1
8	52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
9	57	1	2	150	168	0	1	174	0	1.6	2	0	2	1



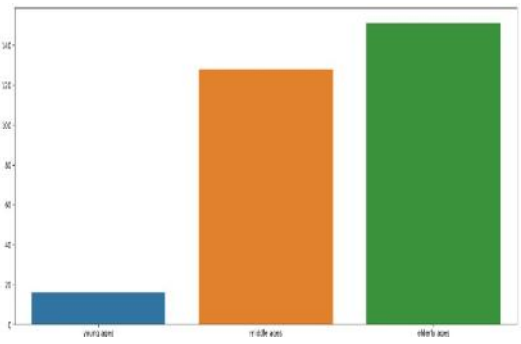
In Above table a box plot (or box-and-whisker plot) method used to describes the distribution of quantitative data in such a way that comparisons across variables or levels of a

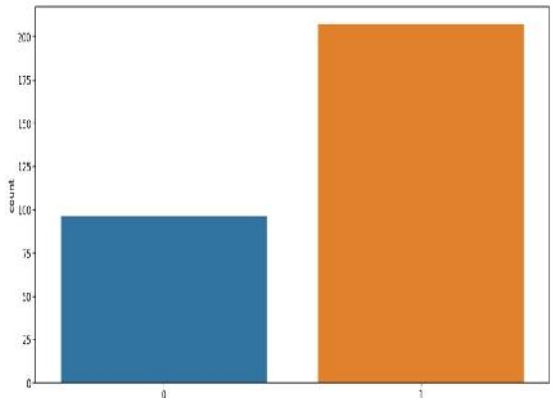
categorical variable are possible. Except for values that are designated "outliers" using an approach based on the inter-quartile range, the box reflects the dataset's quartiles, while the whiskers extend to represent the rest of the distribution.



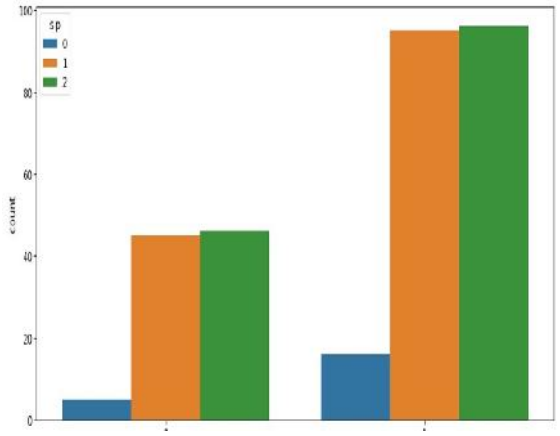


In pie-chart it has three sections in it that are young ages, middle ages and elderly ages have as shown in above figure. In young age group portion shows it has fit and healthy percentage of people which minimizes the healthcare risks at around 95% , while middle age group shows 43.4 % chances of having health issues. Also in elderly ages group there are 51.2 % chances of having heart related or any other health related problems.

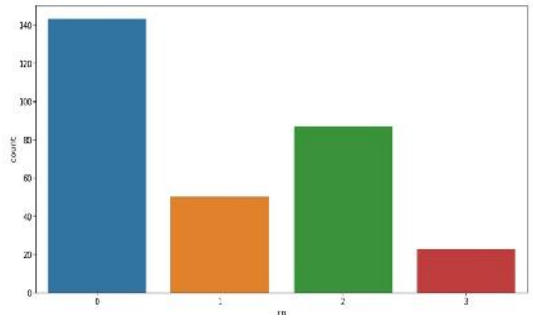




In the above graph, there are three categories of ages as young age, middle age, and third elderly ages. This graph shows that an elderly age group faces lots of problems of their health like diabetics, high blood pressure, cholesterol etc. Heart disease is the leading cause of death for adults over the age of 65.

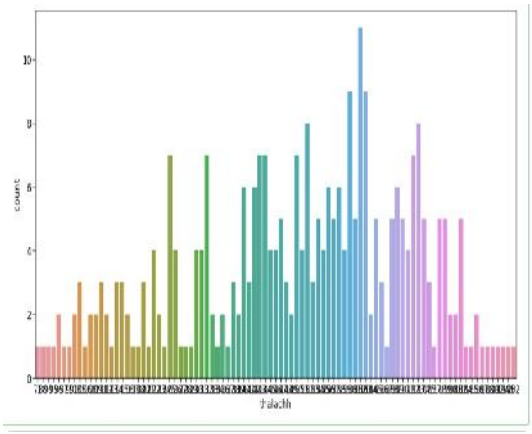


In this graph, young age people has less count as compared to middle age group and elderly age group. Middle age group and elder age group has equal count in both 1 and 2.



In this graph, blue color column represents 44 age group, orange column represents 51 age group, where green column represents 52 age group and red column represents 54

age group. Blue column has maximum count as compared to all remaining three age group columns.



IV. CONCLUSION

Future research objectives include testing chosen algorithms against diverse medical datasets. The studies would be conducted on a greater range of medical data, allowing for a more precise evaluation. It is a good idea to incorporate several algorithms in trials and compare their performance in the medical field. This would result in a new ranking and would contribute in the development of Medical Decision Support Systems by allowing the best relevant algorithms to be chosen. We may also examine different ways not included in this poll to select the optimal one by comparing the advantages and disadvantages of the current one.Various data mining techniques must be used in tandem to enhance sickness prediction accuracy, increase survival rate in crucial death-related scenarios, and so on. To get higher-quality medical data, all necessary steps must be taken to build better medical information systems that provide accurate information about patients' medical histories rather than invoicing invoices. Because high-quality healthcare data is beneficial not only to patients, but also to healthcare organisations and other organizations involved in the healthcare industry. Makes every attempt to bridge the semantic gap in data sharing across distant healthcare database setups in order to discover meaningful trends. These patterns might be very useful in enhancing treatment effectiveness, identifying fraud and abuse, and improving customer relationship management all around the world.

REFERENCES

[1] Sharma, A. (2014). Emerging Applications of Data Mining for Healthcare Management - A Critical Review. 377–382.

[2] Govrdhan, A. (2010). Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. October 2014.

[3] Durairaj, M., &Ranjani, V. (n.d.). Data Mining Applications In Healthcare Sector: A Study. http://www.mfta.ir

[4] ALI, S. I. M., & BUTI, R. H. (2021). Data Mining in Healthcare Sector. MINAR International Journal of Applied Sciences and Technology, 03(02), 87–91. https://doi.org/10.47832/2717-8234.2-3.11

[5] De Vos, A., &Soens, N. (2008). The power of career counseling for enhanced talent and knowledge management. Smart Talent Management: Building Knowledge Assets for Competitive Advantage, 17(8), 119–138. https://doi.org/10.4337/9781848442986.00014

[6] 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. https://doi.org/10.14257/ijbsbt.2013.5.5.25

[7] Jothi, N., Aini, N., Rashid, A., & Husain, W. (2015). Data Mining in Healthcare – A Review. Procedia - Procedia Computer Science, 72, 306–313. https://doi.org/10.1016/j.procs.2015.12.145

[8] Canlas Jr, R. D. (2009). DATA MINING IN HEALTHCARE : Data Mining in Healthcare : Current Applications and Issues. Unpublished Master Thesis, August, 1–10.

[9] Tayade, M. (2014). Role of Data Mining Techniques in Healthcare sector in India Scholars Journal of Applied Medical Sciences (SJAMS) Review Article Role of Data Mining Techniques in Healthcare sector in India. June 2013, 1–4.

[10] De Vos, A., &Soens, N. (2008). The power of career counseling for enhanced talent and knowledge management. Smart Talent Management: Building Knowledge Assets for Competitive Advantage, 17(8), 119–138. https://doi.org/10.4337/9781848442986.00014