Data Analytics in Healthcare Sector

S. Suvetha, Dr. C. Immaculate Mary

[1]Research Scholar, [2] Associate Professor and Head,
Department of Computer Science,
Sri Sarada College for Women(A), Salem, India

Abstract: Data science is currently one of the most emerging fields. Data science is a domain that deals with vast amount of data with a combination of Mathematics, Machine Learning, Statistics, Data Analytics and Artificial Intelligence which helps to extract hidden insight and meaningful information from the data. These derived insights can be used for decision making and strategic planning in various situations. The applications are emerging in many sectors like Healthcare, Information Technology, Media, Education, Entertainment, Banking, e-commerce and financial services. The article reviews and discusses about the process of data cleaning, data preparation and data analysis used in healthcare applications. The article also briefs about advantages and disadvantages of data analytics in healthcare sector. Data science in health care provides real insights and helps in making decision and Data driven Decision Making helps and assist many individuals to improve and adapt to healthy life style.

Index Terms **-: Data Science, Machine Learning, Data Analytics, Healthcare.**

I. INTRODUCTION

In recent times the applications of Data Science are increasing predominantly in various sectors like education, finance, banking, e-commerce, IT, entertainment, healthcare and in several other areas. Recent research says that human body generates 2.5 quintillion bytes of data per day. An Individual data includes all the human activities as well as health records like stress level, blood level, oxygen level etc., to handle all the numerous data we were in need of technology there comes the importance of "Data Science". The healthcare sector holds the upper hand and hardly in need of data science. Even from remote locations it helps doctors to monitor patients records, which is collected using several sources and also helps to detect and diagnose illness at an early stage with the help of machine learning. There are numerous factors that make data science essential in healthcare. In the health industry, human-derived data are highly in demand. The collected data from the valid and proper channel helps to improve the quality of healthcare and it is also used for many health insurance companies, pharmaceutical companies and other organizations.

II. LIFE CYCLE OF DATA ANALYTICS

Data analytics is a process to extract meaningful insights from a raw data. Data is the most valuable resource in today's environment because of the rapid growth and generation of data across all domains. The life cycle will present the overall system involved in data process. The data analytics life cycle binds the process of data generation, collection, processing, and analyzing to extract useful objectives. This life cycle of data analytics will give proper guidance and strategies on how to extract information's from raw data and make it further beneficial [1]. It also helps to make a successful implementation of a model. The Fig.1 represents the step involved in life cycle of Data Analytics.
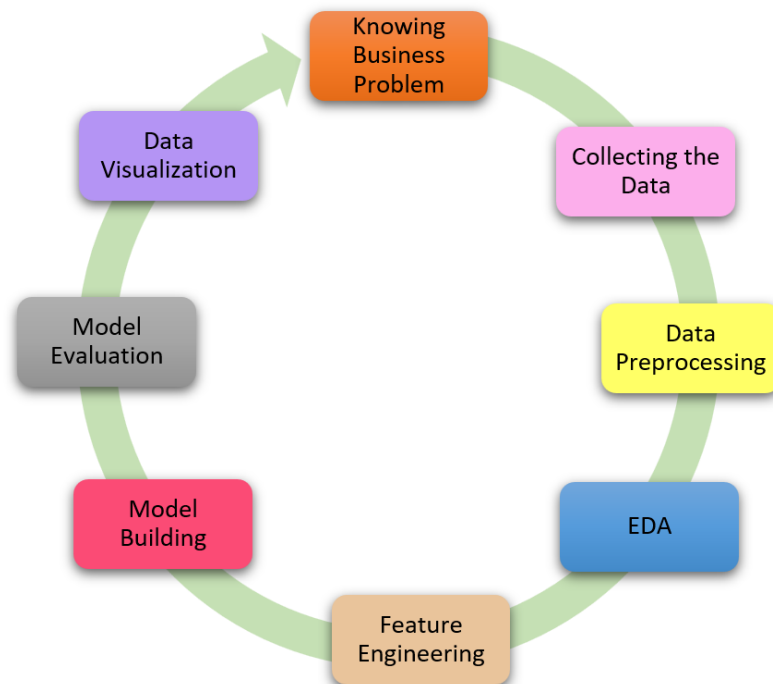
Fig.1: Life Cycle of Data Analytics

The above-depicted Life Cycle will be used by data analysts to advance or reverse their analysis [2]. They will be able to determine, with the assistance of new information, whether they should continue with the same type of work or abandon it and start over [3]. The data analytics life cycle will be used to guide the entire procedure.

2.1 Knowing business Problem

The primary objective of this initial phase is to conduct assessments and evaluations in order to formulate a fundamental hypothesis for resolving any issues or problems in the business.

2.2 Collecting the data

Each and every day there is generation of new data, not all data is collected or utilized. The analyst has to decide what data should be gathered and how to make best use of it. There are many ways to collect the data. using survey, forms and interviews are few methods.

2.3 Data Preprocessing

Once the data is collected from various sources it cannot be directly fed in to the model because the data contain many irrelevant values and null values. So collected data has to undergo for pre-processing. Then the unnecessary null values will be removed and then it can be fed into machine learning model.

2.4 EDA

The term EDA is abbreviated as "Exploratory Data Analysis". It is widely used to Summarise and examine the data. It will also help to recommend the hypothesis and define the pattern between the data.

2.5 Feature Engineering

Feature selection and Feature Extraction is two things which comes under Feature Engineering. In Feature selection we will try to figure our most required and relevant feature among all other variables. In Feature Extraction we will create new feature by combining or transforming the actual attributes.

2.6 Model Building

After discovering the pattern and understandings about the data, now we have to select the optimal model (i.e., algorithm) that suits the dataset. Not all algorithm will be suitable for all kind of data. There are few limitations on deploying the model. The concept of Machine Learning will be discussed in further sections.

2.7 Model Evaluation

After making the model, the new set of data have to fed in to the model to evaluate the accuracy of the model. Then only we can get to know how accurate the developed model is.

2.8 Data Visualization

From the insights gained from the above-mentioned steps, now we can able to consolidate the results and we can analyse whether the model prediction is successful or not. With the help of data visualization, the fetched results will be displayed and presented.

III. LIFE CYCLE OF MACHINE LEARNING

In the real world we the human learn many things easily through our past experiences. In ancient days computer will follow our instructions to perform executions. But in today's world the computer which have the capacity of learning many new things from the old experiences like the way humans does. So here is the role of Machine Learning comes in. Machine Learning is also called as the subset of Artificial Intelligence which mainly deals with the development of algorithms, it will allow the machines to learn from past experiences on their own. Machine Learning (ML) is broadly classified into four types they are illustrated and explained in detail.
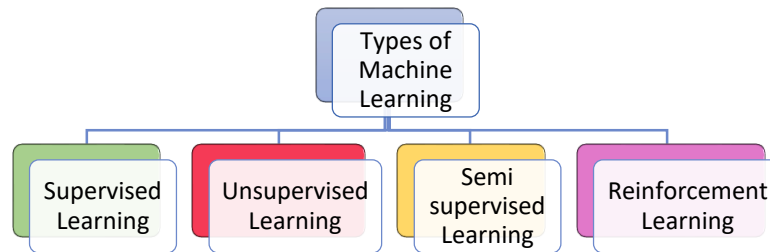
```
                    Types of
                    Machine
                    Learning
    ┌──────────┬──────────┴─────┬──────────────┐
 Supervised  Unsupervised    Semi         Reinforcement
 Learning    Learning        supervised   Learning
                             Learning
```

Fig.2: Types of Machine Learning

3.1 Supervised Learning:

As the name denoted the learning is done with the help of supervisor. Here the term supervisor denotes the labelled values. In supervised learning algorithm the dataset will have labelled attributes. After training the model it will predict the output. The main aim of supervised learning is to map the input variable to the exact output variable.

3.2 Unsupervised Learning:

As the name denoted the learning is done without the help of supervisor. Un supervised learning is direct opposite to supervised learning. There will be no labelled attribute in the data set. only values will be present. The main purpose of this kind of learning is to group the data according to the similarities.

3.3 Semi Supervised Learning:

This kind of learning lies in between supervised learning and unsupervised learning. Here the dataset will have both labelled values and unlabeled values. Semi supervised learning is introduced to overcome the disadvantages of supervised learning and un supervised learning.

3.4 Reinforcement Learning:

Feedback based process is followed in reinforcement learning. Here there will be Artificial Intelligence agent which automatically learns from its own experience. Based on trail and error method it will work. If job get done right reward will be given. If not, the reward will be detected. Based on this kind of process learning will happen. In this there is no labelled data, it learns from its experience only.

There are many machine learning algorithms which are classified among above mentioned learning methods.

- Supervised learning – Classification and Regression.
- Unsupervised Learning – Clustering and Association.

Few popular classification algorithms are Random Forest Algorithm, Decision Tree Algorithm, Logistic Regression Algorithm, Support Vector Machine Algorithm are used to classify the data.

Few popular Regression algorithms are Simple Linear Regression Algorithm, Multivariate Regression Algorithm, Decision Tree Algorithm, Lasso Regression are used to Predict continuous output. Few popular Clustering algorithms are K-Means, DBSCAN Algorithm, Principal Component Analysis used to group data according to similarities.

A cyclical procedure for building an effective machine learning algorithm is the Machine Learning (ML) life cycle.
ML uses multiple processing layers to compose a computational model representing multiple abstraction levels for processed data [7]. The life cycle's primary objective is to resolve the issue or problem. The life cycle of machine learning consists of six major steps, which are listed in the diagram:
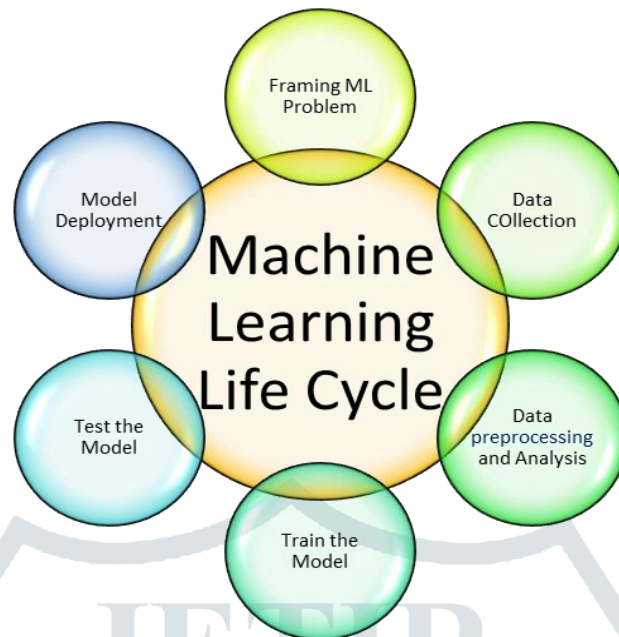
Fig.3: Life Cycle of Machine Learning

A. Framing ML Problem:

The main thing in the entire process is to understand the actual problem. Unless we know the problem, we cannot able to define valid results. Here we will be creating the machine learning algorithm which is called as model and this model will get trained with dataset. So, we require data in this life cycle.

B. Data Collection:

After identifying the actual problem, next step is collect the relevant data from various sources like mobile applications, surveys, existing records etc., The efficiency of the output is completely depends upon the quality and quantity of data collected.

C. Data Preprocessing and Data Analysis:

After collecting and integrating the data from various sources, it has to under go process of cleaning and preprocessing to make use of data in further stages. The preprocessing is done to remove the null values, missing values, duplicate values, noisy data from the data set because it does not serve any purpose.

Later converting raw data into usable format, it will under go processing of data analysis. The main purpose of data analysis is to understand and analyze the data by implementing suitable Machine Learning algorithm like clustering, classification, Regression, Association etc.,

D. Train the Model:

The entire dataset will be generally splitted in to 70:30 ratio. The 70% of the data set will be used for training the model. Based on this data, the developed model will get trained and makes itself ready for predicting the new values.

E. Test the Model:

Remaining 30% of the data will under go for prediction through the developed model. If model performs good enough the accuracy score will be higher else it will be lesser. Through the results we can validate the performance of the developed model.

F. Model Deployment:

If developed model fails to get good accuracy, then reverse process should be done. Again, the data have to be processed and fed to the model after making valid changes. The entire life cycle repeats. When it attains better accuracy results, the model will be deployed in the real - world.

4   DATA ANALYTICS IN HEALTHCARE SECTOR

There are large amount of data is generated from healthcare sector like patient information, their symptoms are collected and stored. There is a need of analyst who plays through the data and get some valuable insights. Now a days data analytics skill is most required to diagnose the illness accurately and helps to save many lives [4].

The steps involved in Data Analytics is listed below:

- Gathering patient data/information.
- Categorizing and structuring the data.
- Implementing the techniques of Data Analytics.
- Deploying a predictive model.

The importance of data analytics in healthcare is clearly exemplifies in the recent times during the Covid-19 pandemic. Only based on collected samples the analyst predicts the next region to be affected and tracks down the symptoms. The whole awareness is taken only because of predictive analytics. In case of covid 19, the model learns from data, collected from affected people and finds new insights like how fast it spreads and cause of spread and additional symptoms are predicted [5]. Machine Learning algorithms corelates and associates the common features like symptoms, habits and sense useful validations.

Predictive analytics plays a vital role in the health care sector as it is data driven approach which completely focuses on prevention of many ailments [6]. It also helps to keep track of patients records and helps to improve their wellbeing. If an illness is predicted before it will be much easier for the doctors to start an early treatment that helps to reduce risk of getting worse.

In today's world the smart watches serve many people and helps them to lead a heathier life. It acts as a personal assistance, and reminds person to take a walk if he/she sits for a long period. Reminds to drink water, if they haven't consumed water for hours. And keeps track of blood pressure level, oxygen level etc., If the analytics is carried over in an efficient way it helps to save many lives.

5   APPLICATIONS OF DATA ANALYTICS IN HEALTHCARE SECTOR

The healthcare sector holds many applications of data analytics. It has been used in various places like discovering new illness, monitoring patient health, keep track of patients records and many more. Few of the applications are taken for consideration and it's use cases are explained below.
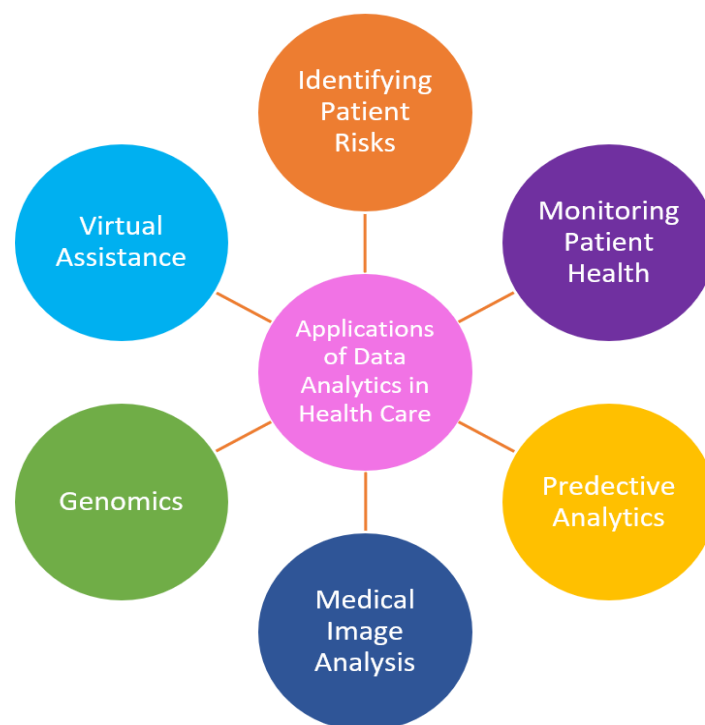

Fig.4: Applications of Data Analytics in Healthcare

5.1   Identifying Patient Risks:

The data of a patient is keep recorded through various sources and it helps to identify who are nearer to the critical condition. If a person Blood Pressure Level goes high it means he is in need of medical emergency. If BP level is recorded and monitored regularly, we can take care of patient and do necessary precautions so it helps to identify the patients who are at risk.

5.2 Monitoring Patient Health:

Data Science plays a vibrant role with help of IoT (Internet of Things). Most of the wearable devices are embedded with IoT [8], which helps to track heartbeat, temperature of the patient. The collected data will be analysed with tools and techniques. The doctor can have a look of patient health condition through remotly.

5.3 Predictive Analytics:

With the fact of previous collected data, the developed model can help to predict what could happen in the future. Based on symptoms of many cases it helps to give precautionary measures and reduces of risk. The condition of patient will get worse if the details are not collected properly. The use of predictive analytics is clearly summarized in section IV.

5.4 Medical Image Analysis:

Analyzing the medical image is one of the most common application of data analytics in healthcare sector. The Data Science recognize the scanned images and finds out the flaws, which helps the doctor to give an exact treatment. The image type can be an X-ray, MRI scan, CT scan etc., once the image is analyzed thoroughly it aids valuable insights and that help doctors to give better treatments.

5.5 Genomics:

Genomics is one of the fascinating research areas in field of medical science. The examination of genes and DNA's is the study of genomics. It helps in finding the characteristics and irregularities in the DNA and corelates the disease, symptoms and affected person's health condition.

5.6 Virtual Assistance:

Data Science is at most used in Virtual Assistance. Through the help of Artificial Intelligence, it gives personalized experience to the patients. The patient has fed the symptoms and the model will predict the cause and suggest some medical remedies to overcome the same. Virtual assistance of doctors mostly helpful in psychological distress and mental health.

6 ADVANTAGES OF DATA ANALYTICS

- It helps to keep track of numerous records generated from hospitals [9].
- It helps to reduce the risk of treatment failure.
- It supports to provide proper treatment on time which could result in saving many lives.
- With the help of virtual assistance, patient need not require to visit the doctor in person. It saves time.
- It helps patients to detect the problem in early stages.
- Because of predictive analytics, it aids to predicts a patient's medical condition and advises to take precautionary measures.
- It helps to forecasts the forthcoming medical emergencies of a patient.

7 LIMITATIONS OF DATA ANALYTICS

- The privacy and security of patient data has to be ensured [10].
- Domain knowledge and technology knowledge is required while making decisions.
- Arbitrary data may mislead the result and cause severe impact on humans.
- The data have to be collected from a valid source. If not, the efficiency of the model will not serve any purpose.

8 CONCLUSION

The performance of the Data Science in the field of Health Care Sector is remarkable. It has been using to treat from minor headaches to tumors. It helps doctors to understand the reason for failures of treatment from the past data and helps to improve the same in the future treatments. The application of Data Science is predominately increasing in all the areas as discussed above. The rapid rise in the development of data Science in the field of Healthcare Sector have both the advantages and its own limitations. Comparatively there are more advantages so, if necessary, steps were taken those drawbacks can also be rectified and we can make at most use of the technology.

REFERENCE

[1] "Data management life cycle final report."

[2] Data Lifecycle Management (DLM) - Spirion.".

[3] "A. M. Cox, W. Wan, T. Tam, A. M. Cox, W. Wan, and T. Tam. (2018) "Acritical analysis of lifecycle models of the research process and research data management,".

[4] "Bytes to Bucks -The Valuation of Data - HealthCare Appraisers.".

[5] H. Khaloufi, K. Abouelmehdi, and A. Beni-hssane. (2018) "Science Direct ScienceDirect Security model for Big Healthcare Data Lifecycle," *Procedia Comput. Sci*., vol. **141**, pp. 294–301.

[6] Koh HC, Tan G.Data mining applications in healthcare. Journal of Healthcare Information Management. 2005; · · (2):64-72

[7] Y. LeCun, Y. Bengio and G. Hinton, "Deep Learning" in Nature,vol. 521., pp. 436-444, 2015.

[8] Sri Venkat Gunturi Subrahmanya, Dasharathraj K. Shetty, Vathsala Patil, B. M. Zeeshan Hameed, Rahul Paul, Komal Smriti, Nithesh Naik & Bhaskar K. Somani - The role of data science in healthcare advancements: applications, benefits, and future prospects.16 August 2021.

[9] AOCNP, D. (2015) The evolution of the electronic health record. Clin J Oncol Nurs 19(2):153.

[10] hvachko K, Kuang H, Radia S, Chansler R (2010) The hadoop distributed file system. In 2010 IEEE 26th symposium on mass storage systems and technologies (MSST) (pp. 1–10). Ieee https://doi.org/10.1109/MSST.2010.5496972.