Analytics for Hospital Health-care Data

DATE:	**03/09/2022**
TEAM ID:	**PNT2022TMID48286**
PROJECT NAME:	**Analytics for Hospitals Health-Care Data**
MAX MARK:	**2 MARK**

Amrish R, Kannadasan P, Ramar R, Thoondimuthu R

ABSTRACT:

Accurately predict the Length of Stay for each patient by case basis so that the Hospitals can use this information for optimal resource allocation and better functioning. While healthcare management has various use cases for using data science, patient length of stay is one critical parameter to observe and predict if one wants to improve the efficiency of the healthcare management in a hospital. This parameter helps hospitals to identify patients of high LOS-risk (patients who will stay longer) at the time of admission. Once identified, patients with high LOS risk can have their treatment plan optimized to minimize LOS and lower the chance of staff /visitor infection. Also, prior knowledge of LOS can aid in logistics such as room and bed allocation planning. Suppose you have been hired as Data Scientist of Health Man – a not for profit organization dedicated to manage the functioning of Hospitals in a professional and optimal manner.

Literature survey

Conley et al (2008) Predictive analytics supports healthcare sectors to achieve a high level of effective overall care and preventive care, as predictive systems' results allow treatments and actions to be taken when all the risks arerecognized in early stages, which aids for minimizing costs. Further more. Obenshain (2004) said that patients can also work and support medical care by following up and updating their medical status, so they can get the necessary

treatment at the right time. The technology era has added significant value to the healthcare decision support system, Cannon &Tanner(2007)since decision making systems in healthcare care sectors can be enhanced by focusing on patient diagnoses, behavior, and prevention in order to reach a high level of care and improve healthcare economics.McHorney (2009) has added that healthcare analytics is not solely regarding technology and the knowledge however; it is also regards how much individuals are attached to and familiar with medical care systems and their personal skills such as ability to learn and adopt such systems in their life, as different people have different attitudes and reasons for not accepting such technologiesBrownstein & Wicks (2010) patients can share some information with other patients, so they increase their knowledge, background and awareness in the healthcare analytics sectors regarding their conditions. Finally, patients who share their symptoms, diagnoses and results with others can gain benefit from the ability to understand their health conditions by comparing them with other patients.Russom (2011) analysed and stored in order to produce useful and high quality information and knowledge. This term also includes the way of how this data is gathered, filtering and preparation of the data for use and finally the processing of data to support data analytics and predictive modelling. Turner (2011) agreed that social media and internet applications have a big influence on collecting patient information through filling and completing some online forms in order to keep track of their state of health, as well as to provide the suitable advice and treatment when needed.Miron et al (2011) believed that whatever and how much our patients are educated and skilled to provide us with the data we expect, medical professionals still highly need to test and clarify this data to consider it and keep it on record. Also, he added that once when the data has been tested and clarified, we then need to find out how to change an individual's behavior starting with parents and guardians who are responsible for raising their children.Swan (2012) was discussing the same point when he identified the term "citizen science", where nonprofessional and educated individuals are skilled enough to conduct and support healthcare analytics system. Accordingly, this will require organizations to train individuals how to follow up and track their health information, as well as self-monitoring. Loginov et al (2012) stated that by retrieving and reviewing past patient details, information and diagnoses from the databases, predictive methods can take a place through forecasting, reducing time and costs.Jacob (2012) Parkland hospital in Dallas, Texas has launched a predictive system which scans all patient's details and information to identify potentials risks and outcomes. As a result, the hospital has saved more than half a million dollars, especially in heart failure and disease predictions in terms of performing patients' monitoring and avoiding future complications .Kim (2013) sometimes being motivated for change and in understanding of information are

not enough. Furthermore, patients should identify the risks and detect where to change, for instance; some patients know that they have a high level of blood pressure but they don't know how to deal with it and control it.LexisNexis(2015) Healthcare prediction is another data analytics method focusing on reducing future medical costs. Predictive technique uses patient medical history to evaluate all the potential health risks and predict a future medical treatment in advance.Lamont (2010) says electronic clinical records to analyze diagnosis and confirm outcomes in order to provide the correct treatment for the right patient at the right time. Moreover, Imamura et al (2007) has found that the associateion diagnostic approach can effect efficiently in extracting desirable information from huge databases.Bertsimas et al (2008) Says Medical care systems have focused on increasing healthcare analytics performance as well as minimizing the cost by simplify unstructured clinical record and reducing irregular information. Consequently, large quantities of information then will be managed and controlled smoothly and efficiently.Andreu-Perez et al (2015) Veracity is crucial for Big Data analytics. Personal health records (PHRs) may contain abbreviations, typographical errors, and cryptic notes. Ambulatory measurements are possibly completed under uncontrolled and less reliable environments compared with clinical data which is collected by trained practitioners in a clinical setting. Using spontaneous unmanaged data from social media may result in inaccurate predictions. In addition, data sources are sometimes biased.Sacristán and Dilla (2015) 'Noise' data is a massive problem especially when it grows fast. Databases with various degrees of completeness and quality lead to heterogeneous results, which increase the possibility of false discoveries and 'biased fact-finding excursions'. Low data quality and biases due to the absence of randomization are two major problems. Efforts in increasing the value of big data are often made through linking different databases and analyzing all existing and related data.Farid et al (2016) Data pre-processing is a process of transforming raw data into an understandable format that often includes: 1) data cleaning, 2) data integration, 3) data transformation, 4) data reduction, and 5) data discretization. The pre-processing is an important step for Big Data analytics.Simonsen et al (2016). Systems relying on big data streams have been developed, which include patient-level hospital discharge records, electronic death certificates, and medical claims data that use International Classification of Diseases (ICD) coding.Salavati et al (2017) Surveillance tactics using big data streams from crowd sourcing, social media, and Internet search queries have been proposed Big Data technologies like NoSQL databases have been used in processing healthcare information, while some features like local access and rational relationship between logical and physical data distribution are important to improve the performance of parallel processing in distributed

databases.A Big Data–driven approach and process was proposed that incorporates both clinical and molecular information. Candidate biomarkers and therapeutic targets/drugs are fists identified in the approach. Subsequent clinical or preclinical validation is completed by the cross-species analysis; therefore, the required costs and time in biomarker/therapeutic development are reduced (Wooden et al., 2017).Istephan and Siadat (2015) A clinical data warehouse was created for structured data; a set of modules were also built for analyzing unstructured content. The research was conducted to build an initial implementation of a framework within a big data paradigm. The framework runs the modules in a Hadoop cluster and the distributed computing capability of Big Data was used A Hadoop-based architecture was developed to manage Twitter health big data.Cunha et al (2015) says Analyzing tweets in healthcare has the potential to change the way people and healthcare providers use advanced technologies to achieve new clinical insights. Vanathi and Khadir (2017) Open sources such as Hadoop, Kafka, Apache Storm, and NoSQL Cassandra have been used in Big Data analytics. There are a set of general primitives in Apache Storm for computing real-time big data.Research on attribute reduction has been done using MapReduce based on the Rough Set Theory (RST). The procedures include 1) use parallel large-scale rough set methods for feature acquisition and implement them on MapReduce runtime systems such as Twister, Phoenix and Hadoop to obtain features from big datasets through data mining; 2) use the framework structure of < key, value> pair to accelerate the computation of equivalence classes and attribute significance; parallelize traditional attribute reduction process based on MapReduce (Ding et al., 2018). Traditional high-performance computing (HPC) is computation (CPU) oriented with intensive computing through internal (supercomputing) or external high-performance networking (cluster or grid computing), while Hadoop-enhanced computing is intensive computing for largescale distributed data through internal and external networking. Hadoop-based Big Data has three advantages: efficiency, reliability, and scalability (Ni et al., 2015). Table 6 (Olaronke and Oluwaseun, 2016) shows a comparison of tools used for analyzing big data in the healthcare systemÖzdemir and Hekim(2018) Industry 4.0 is a strategic plan in manufacturing and custom manufacturing of medical devices and drugs are included in Industry 4.0. Precision medicine is a kind of Big Data application in health, which benefit from multi-omics, IoT, Industry 4.0, etc. Industry 5.0 has been proposed which make sense of Big Data with artificial intelligence, IoT, and next-generation technology policy.An intelligent healthcare framework has been developed based on IoT technology to provide ubiquitous healthcare for a person during his/her workout sessions. An artificial neural network model was used to predict the person’s health related vulnerability using Bayesian belief

network classifier. Data management, model development, visualization, and business models have been listed as four key areas of Big Data analytics Verma, and Sood (2018). Some data mining methods for complex EHR big data are summarized in Table 7 (Wu et al., 2017).Integration of physiological data with high-throughput "-omics" techniques for clinical recommendations is also a challenge. The continuous increase in available genomic data and related effects of annotation of genes and errors from analytical practice and experiment have made the analysis of functional effects using high-throughput sequencing methods a challenging work Belle et al., 2015. The issue on consent to using healthcare data such as genetic data has been a concern. Creating databases based on large and national population for future research with ethics approval and governance has led to academic debates on legality. There are even arguments on that Big Data is useful to improve healthcare systems (Knoppers and Thorogood, 2017). The following are general challenges of Big Data in healthcare (Mathew and Pillai, 2015).

Tools:

Tools	Description
Hadoop Distributed File System(HDFS)	HDFS is a sub-project of the Apache Hadoop project. This Apache Software base project is designed to provide a fault- tolerant file system designed to run on service hardware.
MapReduce	Map Reduce is an associated implementation and programming model for processing and generate large data sets through a parallel, distributed algorithm on a cluster.
Hive	Hive it permits SQL programmers todevelop Hive Query Language (HQL) statement similar to typical SQL statement[1].
Zookeeper	ZooKeeper is an open source Apache project that is provides a centralized infrastructure and services that allow synchronization across a cluster.
HBase	HBase is an oriented database management system that works on top of HDFS. It is well matched for sparse data sets, which are common in many of bigdata use cases. It uses a non-SQL approach.
Cassandra	Cassandra is a distributed database system. And it is chosen as a top-level project modeled to handle big data shared across many utility servers.
Mahout	Mahout is yet another Apache project whose objective is to create free applications to distributed and scalable mechanism knowledge of algorithms that support by big data analytics in the Hadoop platform.