Analysis of Health Care Data By Data Analytic Tools

[1]Aditya Gupta, [2]Suhani Jain
[1]Electronics and communication,[2]Computer Science
[1]Medi-Caps Institute of Technology and Management,Indore,[2]Mody University,Laxmangarh

---

Abstract - Health care is one of the biggest sector in India and it plays major role in growth of the country.The amount of data generated is increasing day by day and also need to get meaningful insight out of it also.The health care sector is also generating enormous amount of data each year.Thus the data generated by health care sector can be used to get meaning full insight about doctors patients,diseases,sex ratio,health ratio which helps in improving health care System .The main purpose of this paper is the data analysis and visualization of health care data by using big data and analytics tool.

IndexTerms - Data analytics,Health care,Big data.

I. Introduction

Big data value has been increased in the recent years especially in health care industry.Initially enormous amount of data generated by the health care industry was stored as hard copy. This data has the ability to support a wide range of health care and medical functions. The digitization of such data is called Big Data. The data is related to patient health care and well-being makes up big data. A 2011 McKinsey report estimated that the health care industry could potentially realize $300 billion in annual value by leveraging big data.Big data in health care opens new way of treatment and also at the same time provides us way for improving this system.

II. Big Data

Big data is one of the most cardinal factor in data analysis. It is the major element which is used in the data analysis.Big Data is a collection of large datasets that cannot be processed using traditional computing techniques. Big Data involves various tools, strategy, techniques,testing, challenges, and advantages. Big Data is a vast field. Day by day there is some new evaluation is taking place. There are many future aspects of Big Data,and there are many field in Big Data in which we need to work. Here we will discuss the Big Data strategy, some technique, and the Big Data testing – various types of testing, challenges, future aspects and the advantages.process to be followed, so as to maintain as well as manage the important characteristics of Big Data like Volume, i.e. data size, Velocity, i.e. speed of change and Variety of data sources.One of the best methods which can be used for testing Big Data apps is testing with automation.

The Big Data has 3 vs- volume, variety and velocity.
Volume - Here volume of the Data defines and tell us that the data has been collected from various sources like business, transport, media etc. And thus a large amount of data has been collected and by that the volume increases.
Velocity - The velocity is basically the speed by which the data are transferred from one place to another. At what speed the data is being share. Is it in a real time.The basic idea behind the velocity is that the maximum data should be transferred with the maximum speed.
Variety – There are different types of data like structured, unstructured- in that there are many others like video audio, text, messaging, and email. These all comes under the section variety.

III. Big Data Challenges

In today's world, Big Data proves to be beneficial and positive for all of us.But there is also some certain challenges which the process of Big Data faces. It's not just a small process, but actually it has to take care of certain things. There are various challenges which are as follows-

- Security of the dataset.
- Proper maintaining and implementation of the data
- Control over the data
- Listing of the data in the proper arrangement
- Dealing with the data growth
- Validating the data
- Checking of data after the regular interval of time
- Deleting and modifying the data
- Organize the data
- Proper output

•

IV. TOOLS AVAILABLE FOR DATA ANALYTICS

1.**Tableau :**

Tableau Public is a free software that connects any data source be it corporate Data Warehouse, Microsoft Excel or web-based data, and creates data visualizations, maps, dashboards etc. with real-time updates presenting on web. They can also be shared through social media or with the client. It allows the access to download the file in different formats. If you want to see the power of tableau, then we must have very good data source. Tableau's Big Data capabilities makes them important and one can analyze and visualize data better than any other data visualization software in the market.

2. SAS:

Sas is a programming environment and language for data manipulation and a leader in analytics, developed by the SAS Institute in 1966 and further developed in 1980's and 1990's. SAS is easily accessible, manageable and can analyze data from any sources. SAS introduced a large set of products in 2011 for customer intelligence and numerous SAS modules for web, social media and marketing analytics that is widely used for profiling customers and prospects. It can also predict their behaviors, manage, and optimize communications.

3. Apache Spark

The University of California, Berkeley' s AMP Lab, developed Apache in 2009. Apache Spark is a fast large-scale data processing engine and executes applications in Hadoop clusters 100 times faster in memory and 10 times faster on disk. Spark is built on data science and its concept makes data science effortless. Spark is also popular for data pipelines and machine learning models development.Spark also includes a library – MLlib, that provides a progressive set of machine algorithms for repetitive data science techniques like Classification, Regression, Collaborative Filtering, Clustering, etc.

4. Excel

Excel is a basic, popular and widely used analytical tool almost in all industries. Whether you are an expert in Sas, R or Tableau, you will still need to use Excel. Excel becomes important when there is a requirement of analytics on the client's internal data. It analyzes the complex task that summarizes the data with a preview of pivot tables that helps in filtering the data as per client requirement. Excel has the advance business analytics option which helps in modelling capabilities which have prebuilt options like automatic relationship detection, a creation of DAX measures and time grouping.

5. RapidMiner:

RapidMiner is a powerful integrated data science platform developed by the same company that performs predictive analysis and other advanced analytics like data mining, text analytics, machine learning and visual analytics without any programming. RapidMiner can incorporate with any data source types, including Access, Excel, Microsoft SQL, Teradata, Oracle, Sybase, IBM DB2, Ingres, MySQL, IBM SPSS, Dbase etc. The tool is very powerful that can generate analytics based on real-life data transformation settings, i.e. you can control the formats and data sets for predictive analysis.

6. KNIME

KNIME Developed in January 2004 by a team of software engineers at University of Konstanz. KNIME is leading open source, reporting, and integrated analytics tools that allow you to analyze and model the data through visual programming, it integrates various components for data mining and machine learning via its modular data-pipelining concept.

7. QlikView

QlikView has many unique features like patented technology and has in-memory data processing, which executes the result very fast to the end users and stores the data in the report itself. Data association in QlikView is automatically maintained and can be compressed to almost 10% from its original size. Data relationship is visualized using colors – a specific color is given to related data and another color for non-related data.

8. Splunk:

Splunk is a tool that analyzes and search the machine-generated data. Splunk pulls all text-based log data and provides a simple way to search through it, a user can pull in all kind of data, and perform all sort of interesting statistical analysis on it, and present it in different formats.

V.DATA ANALYSIS AND VISUALIZATION

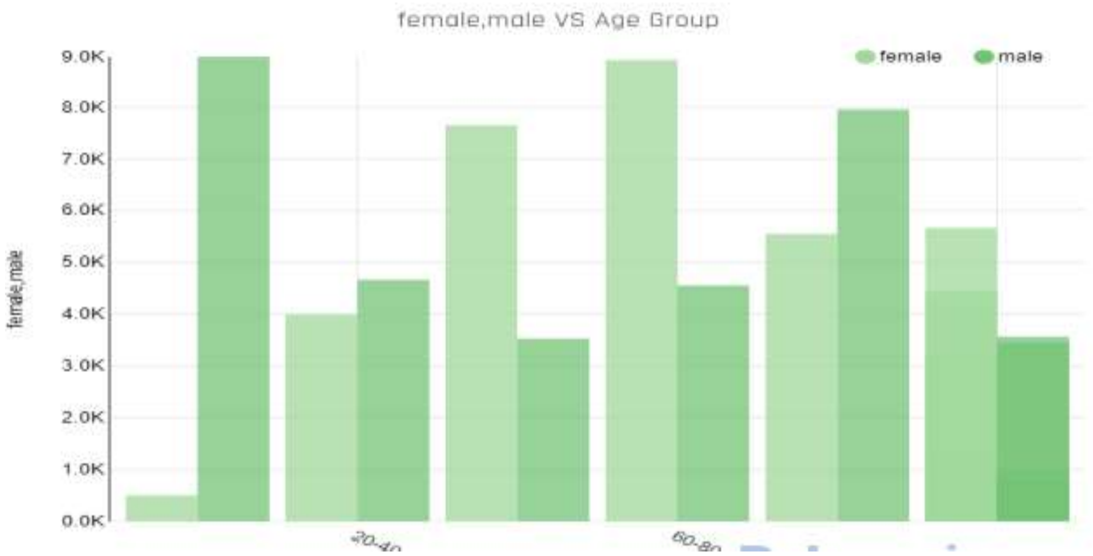

Figure 1 Female,Male VS Age Group

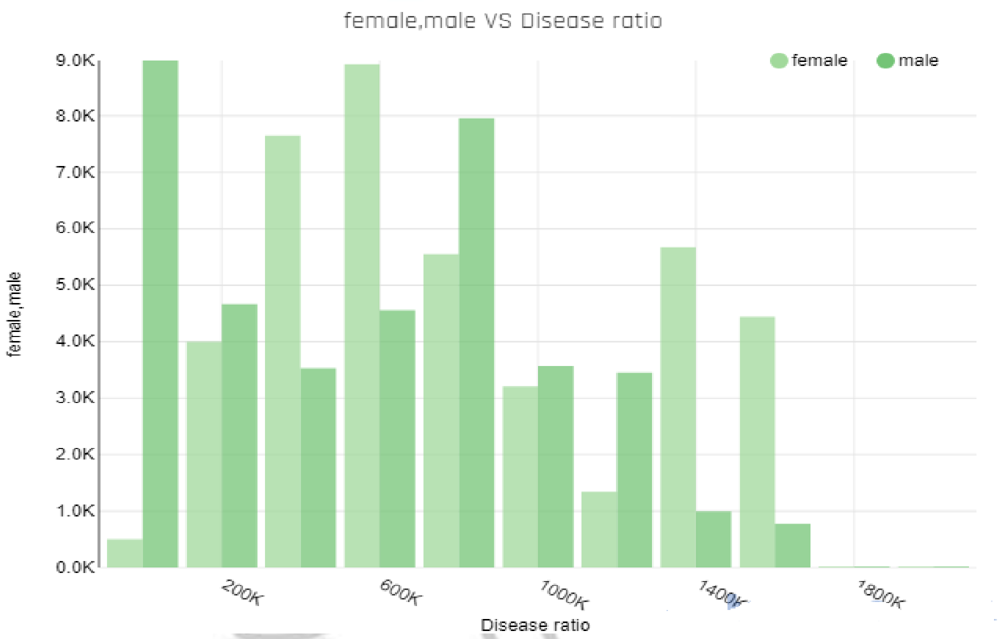

Figure 2 Female,Male Vs Disease Ratio

Based on the acquired data sets and performed the data analysis and displayed the result in the form of the bar graph. The bar graph is constructed with the help of the data analytics tools like tableau and datacopia. These are the some data visualization tools by which data can be presented to make it more understandable. Result obtained from the acquired data sets is constructed in form of graphs and the graph is between health ratio of the female and male in the different areas of the India. The ratio of female and male is not equal in the health care sector. The health problem is generally more in the female as compared to the male. Comparison is also made of the age wise ratio of the female and male health as shown in figure 1..The female are facing greater difficulty in the health care. The analysis is based on the last few years. So from the graph we can easily analyse any field and any data within few time. This gives the more flexible, more accurate result and more advanced result. The use of the data analytics is very useful in the health care sector for the increasing the health ratio of the nation.

VI References

[1] https://www.kaggle.com/rajanand/key-indicators-of-annual-health-survey/home for health analytics.
[2] https://www.kaggle.com/rajanand/suicides-in-india
[3] https://data.gov.in/resources/annual-health-survey-mortality-schedule-description-variablesparameters
[4] https://data.gov.in/catalog/foreign-tourist-arrivals-india-top-15-source-countries if u r interested in travel.- this can be done.
[5] https://www.dataquest.io/blog/free-datasets-for-projects
[6] https://www.springboard.com/blog/free-public-data-sets-data-science-project