SYMPTOMS BASED DISEASE PREDICTION USING BIGDATA ANALYTICS

Dr Balajee Maram¹, Dr K Suresh Kumar², Dr Veerraju Gampala³

¹Associate Professor, Dept of CSE, GMR Institute of Technology, Rajam, Andhra Pradesh
balajee.m@gmrit.edu.in

²Associate Professor, Department of Information Technology,
Saveetha Engineering College, Thandalam, Chennai
sureshkumar@saveetha.ac.in

³Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur, Andhra Pradesh 522502, India.
veerraju5a8@gmail.com

ABSTRACT

The data is being collected from various sources in the world which belongs to Healthcare sector and is processed called Healthcare data management. The huge amount of healthcare data must be processed by efficient tools and methods to create value. In technical market many BI are available to handle structured data only. But the unstructured data is also being generated which can be used to give valuable insights to improve the quality in healthcare. For understanding the patient needs, there is a need to collect structured and unstructured data from various stakeholders. Then the analysts get the whole idea about the patient’s needs based on symptoms and able to give precision driven care and treatment. The final treatment depends on the patient’s present condition and earlier treatment which increases the perfectness in the treatment. This paper the consideration of 400 symptoms and 147 diseases. It analysis the performance of the machine learning algorithms including Decision-tree, Randomforest, Naïve Bayes and the proposed algorithm.

Keywords: Disease, Symptom, Decisiontree, Randomforest, Naïve Bayes

I. INTRODUCTION

Big data shows its significance in every field in the world including healthcare industry. It changes the way to handle the patients and doctors with care. From more number of sample data, can expect more accurate insights for healthcare industry. Like many industries, healthcare industry is a framework which contains heterogeneous sectors are complex to handle with high accuracy, where the patients demanding better care with less price. Day by day, new technologies are being included to the healthcare industry, where the big data analytics plays a vital role for giving effective business insights to the hospitals as well as patients.

In the technical world, data analysis plays an important role in every field in the world where the data volume is so limited. But today, the world is in big data era. The existing statistics says that the data analytics is very important in near future for healthcare industry and it becoming very crucial in clinical, operational and financial sectors.

The collected data can potentially be used by the Govt. and public organizations create or improve policies, procedures, and trainings. Overall, the project has the potential to heighten awareness for the need to give best treatment in any healthcare environment.

Most of the patients are illiterates and those are not familiar precision treatment. So majority of people approaching private health care centers which are not able to store the details of patients and their diseases. So there is a need to organize health camps which educates and sensitize the community. This framework explains about diagnosis and various types of health hazards.

The objectives of the proposed algorithm are as follows:
• To map high-risk areas for disease prevention
• To devise the framework for sharing Electronic Health Records (EHRs) via secure information systems
• To devise dynamic descriptive decision tool for Real-Time Alerting to design security enhanced features
• Telemedicine – It is a process to provide the customized treatment for each patient for avoiding re-admission in hospital again and again.

II. LITERATURE SURVEY

The framework (Prableen Kaur et al., 2018) has been proposed for healthcare system, has four layers. The advantages of this framework are data optimization and data security. It is based on distributed model and enhances the performance of the system by data and storage optimization.

The data processing concept (Sunil Kumar et al., 2019) has been explained. The healthcare data is being generated and coming from various sources in the form of EHRs, genome database, text and imagery unstructured data, clinical reports, sources belongs to Govt. sector, lab reports from medical centres and pharmacies and health insurance companies. This data can be handled by HADOOP framework.

Fang et al. (2016) proposed a framework titled “Health informatics processing pipeline framework” which consists of data capturing, storing, analysing, searching and decision support. It offers dynamic services to the patients through mobile devices and sensor networks. Legaz Garca et al. (2016) proposed a framework based on OWL. It gathers patient data (EHR) and utilized for data exploration.

Sakr and Elgamma (2016) proposed a method that integrates sensors, cloud, IoT and Big data analytics. It is able to handle patient profile analytics, population management etc. But not able to handle complex data sources such as images and streams. Pramanik et al. (2017) proposed a layered framework on healthcare system. This framework yields useful smart system services.

Dencelin and Ramkumar (2016) proposed a framework for analysing big data with the help of Apache Spark. It applies machine learning algorithms using different set of input features and network parameters. D.W. Bates et al.(2014), Big data analytics can help early disease detection, deviation from healthy state and detection of fraud. It also helps in getting accurate predictions, cost-reduction in healthcare maintenance and it provides precision good health.

The framework which contains layers has been proposed for healthcare system by Raghupathi and Raghupathi (2014). The data source layer handles internal and external data sources for healthcare system. The transformation layer for transformation and loading the data. The analytics layer for querying, reporting and processing. Theatrically these concepts are good enough.

After data collection, predictions can be done using machine learning algorithms. For healthcare industry data, two types of prediction algorithms are required. Those algorithms are supervised and un-supervised.

Decision tree is very simple for gaining accurate and fast result, but the tree construction takes lot of time. It assigns class label to each patient for solving various problems. Random-Forest selects the most effective answer. It's appropriate for big knowledge bases and offers correct results by estimating missing data. Naive Bayes is the most simple algorithm that you can apply to your data. As the name suggests, Naive Bayesian classifier is based on Bayes’ theorem with the independence assumptions between predictors.

III. METHODOLOGY

Data-Collection

The data can be collected from various repositories and store the entire data in Hadoop Distributed File System (HDFS). The data can be collected through surveys and questionnaires, focus groups, interviews, and observations and progress tracking.

Sources of Data

Data can be collected from different sources like hospitals, medical practitioners, patient health history, surveys, medical bills etc.
Patient Medical Records: The health history of the patient and diagnostics report can maintained in a single document called medical record or Electronic Health Records (EHRs) (Objective-3). It is readily available to both the patient and hospitals through electronic medical records.

Patient Surveys: This is the process for gathering the data from the various types of patients about their diseases, medical reports, treatment procedure, type of doctor, cost of the treatment, effect of the treatment, billing system etc.

Comments from Individual Patients: Today social networking websites plays a vital role for gathering the opinion from various types of patients in healthcare industry. It gathers the comments from patients informally rather than by prepared questionnaire.

Standardized Clinical Data: The detailed information about each patient can be gathered from clinical and nursing homes, diagnostics centers and health agencies.

This concept was implemented through Python programming language and machine learning algorithms. The list of symptoms and diseases are stored in the form of dataset. The dataset contains 400 symptoms and 147 diseases which belongs various categories of diseases. Usually the user can enter the list of symptoms in the system. Then the system will finds the possible diseases as predictor-1, predictor-2 and predictor-3. The functionality of the proposed system is as follows:

Method of Processing and Analysis

The data for healthcare domain is being generated from various internal and external sources in the world. The data can be gathered and processed in the following:
Web and social media data: The data from the social networking websites like Facebook, Twitter, LinkedIn, health plan websites and various apps.

Machine to machine data: Most of the unstructured data is being generated from sensors, meters and wearable devices.

Big transaction data: Patient join report and discharge reports, medical bills, health care claims, medical images are available in semi-structured and unstructured formats only.

Biometric data: Finger prints, genetics, handwriting, retinal scans, blood pressures, blood sugar, pulse and other personal details of each patient related to his body.

Clinical data: semi-structured and unstructured data such as EMRs, physician’s prescription, email, telemedicine details etc.

After collecting the raw data, it can be stored in a data warehouse. Then the big data analytics process the entire data and all types of data. Then it handles various queries, it generates reports, OLAP and data mining. In big data era, many techniques and methods have been developed for aggregate, manipulate, analyze and visualize the healthcare data.

The healthcare big data can be handled by open-source data processing platform called HADOOP from Apache. Hadoop is based on horizontal scalability and is able to process extremely large amounts of data by large number of clusters of nodes, each node solve some part of the problem, integrates them for the final result. This project can handles the issues related to healthcare industry including ownership, privacy, security and standards.

IV. RESULTS AND DISCUSSION

The framework accepts the username and list of five symptoms. Then it applied Decision-Tree, Random-Forest and Naïve-bayes classifiers on training and testing data. Finally it gives predictions. Similarly the proposed
algorithm also works on training and testing data, finally it gives 3 predictions of diseases based on the given symptoms as shown in the following figures:

The following table Table-1 shows the list of symptoms from the patient:

Table 1: List of symptoms given by the patient

<table>
<thead>
<tr>
<th>S.No.</th>
<th>SYMPTOMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>orthopnea, fatigue, dyspnea on exertion, dyspnea, shortness of breath</td>
</tr>
<tr>
<td>2</td>
<td>drowsiness, sleepy, pain chest, angina pectoris, pressure chest</td>
</tr>
<tr>
<td>3</td>
<td>Wheezing, Cough, Shortness of breath, Chest tightness, Distress respiratory</td>
</tr>
<tr>
<td>4</td>
<td>Hematuria, tumor cell invasion, pain, anosmia, thicken</td>
</tr>
</tbody>
</table>

The above symptoms taken the machine learning classifiers as well as the proposed algorithm. Finally it predicts and shows the possible diseases based on the given symptoms. The final results as shown in the following Table-2.

Table 2: List of predictions given by Decision-Tree, Random-Forest, Naïve-Bayes and proposed algorithm

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Decision Tree</th>
<th>Random Forest</th>
<th>Naïve Bayes</th>
<th>Proposed Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>carcinoma of lung</td>
<td>adenocarcinoma</td>
<td>exanthema</td>
<td>failure heart</td>
</tr>
<tr>
<td></td>
<td>encephalopathy</td>
<td>encephalopathy</td>
<td>encephalopathy</td>
<td>cardiomyopathy</td>
</tr>
<tr>
<td></td>
<td>Exanthema</td>
<td>Sepsis (invertebrate)</td>
<td>Sickle cell anemia</td>
<td>Ischemia</td>
</tr>
<tr>
<td></td>
<td>malignant tumor of colon</td>
<td>pancreatitis</td>
<td>encephalopathy</td>
<td>neoplasm</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>neoplasm metastasis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>coronary artery disease</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Asthma</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Chronic obstructive airway disease</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
After giving the predictions, the accuracy has been calculated and shown in Table-3.

Table-3: Accuracy of Decision-Tree, Random-Forest, Naïve-Bayes and proposed algorithm

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Decision Tree</th>
<th>Random Forest</th>
<th>Naïve Bayes</th>
<th>Proposed Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input#1</td>
<td>0.8911564</td>
<td>0.9047619</td>
<td>0.9047619</td>
<td>0.976</td>
</tr>
<tr>
<td>Input#2</td>
<td>0.9047619</td>
<td>0.9047619</td>
<td>0.9047619</td>
<td>0.984</td>
</tr>
<tr>
<td>Input#3</td>
<td>0.8911564</td>
<td>0.9047619</td>
<td>0.9047619</td>
<td>0.986</td>
</tr>
<tr>
<td>Input#4</td>
<td>0.9047619</td>
<td>0.9047619</td>
<td>0.9047619</td>
<td>0.934</td>
</tr>
</tbody>
</table>

Among Decision-Tree, Random-Forest, Naïve-Bayes and proposed algorithm, the accuracy is best in the proposed algorithm comparatively Decision-Tree, Random-Forest, Naïve-Bayes.

Benefits

The outcome of this paper is one type of software tool only. It collects and maintains very huge amount of data from various sources in the world related to health. It generates various types of reports and insights which gives precision treatment for the patients. The following are the expected benefits from this project:

- Medication is error-free
- Identification of high-risk patients easily
- It reduces hospital visits frequently
- It reduces patient waiting time in hospitals

Limitations and future enhancements

Data Aggregation Challenges: The data related to healthcare system is being generated from various sources in the world like hospitals, administrative offices, Government offices, Private medical practitioners, Laboratories etc. Pulling it together and create a big data set is a challenge.

Polity and Process Challenge: After getting big data set, there is a need to protect it. This project can provide the protection through access control, authentication, encryption and decryption etc. But some part of the data is in the hands of cloud providers which are 3rd party vendors. So privacy and protection may be one of the challenge.

Management and Administration: It is one of the upcoming technology for the staff in healthcare industry. Again there is a need to recruit new IT experts for using HADOOP framework or need to give training to the existing staff. It increases the cost to the hospital authorities.

V. CONCLUSION

Each country taking more care about human health issues in today’s world. Always WHO used to give so many suggestions for preventions of many epidemics or diseases. Today the entire world is giving more importance to identification and prevention of many diseases based on the various symptoms from patients. In Bigdata era, huge amount of data is being generated from various sources in the world. So Bigdata analytics plays an important role for getting predictions in health-care industry. The health-care industry is in a position to predict the disease based on given symptoms and it will provide the suggestions healing of diseases. It reduces the patients to rejoin in the hospitals unnecessary. Physicians also get many suggestions about the good treatment for the patients. It provides the exact treatment for the patients and will provide the exact medicine. Automatically it eliminates the side-effects for the patients. Finally it helps the patients, Doctors, Hospitals.

REFERENCES