International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 11 | Nov 2020 www.irjet.net p-ISSN: 2395-0072

Predicting Healthcare Logistics using Data Science

Aashutosh Awasthi[-1], Sagar Borkar[-2], Naman Bangad[-3], Urvi Shukla[-4], Pratik Dabre[-5]

[1-4]Student, Computer Engineering Department, Thakur College of Engineering & Technology, Mumbai, Maharashtra, India
[5]Student, Computer Engineering Department, Fr Conceicao Rodrigues College of Engineering, Mumbai, Maharashtra, India

---

Abstract*— *The world is moving with a fast speed and in Order to keep up with the whole world we tend to ignore the symptoms of disease which can affect our health to a large extent. Diseases can affect people not only physically, but also mentally, as contracting and living with a disease can alter the affected person's perspective on life. Death thanks to disease is named death by natural causes. Because of improved living conditions and increased access to medications, the proportion of human deaths caused by infectious diseases has trended downwards over the last centuries, giving way to degenerative and lifestyle diseases. The pandemic has made the whole world work from home and the online system will be the new normal. To come up with it we proposed a model which will collect information from the patients and predict the disease along with many additional features.

***Keywords*— Machine learning, data science, COVID-19, SVM, Logistic regression, Decision tree.**

1. INTRODUCTION

The uncertain pause in the growth of the world has led to change the overall system of the world. This pandemic has led the world to move to the new normal which is completely online. Due to the uncertain pandemic, a common problem which was observed was due to the fear of COVID -19 the people are being afraid to visit doctors which led to an increase of many diseases.. To overcome this problem we come up with a model which will predict the diseases based on the symptoms provided by the user along with many additional features as it will also provide the list of clinics and doctors to the user based upon disease and locality ,and many more basic and additional features.

2. OVERVIEW OF EXISITNG SYSTEM

The current or primary system of disease prediction or determination of the cause behind a person's illness is based on a doctor trying to predict it by looking at the symptoms of the patient. In this process the doctor who basically has the medical experience, judges the condition of a patient and then tries to predict the possible disorder and there by suggests a solution in alignment to the initially predicted illness, which could help the patient be healthy again but if this fails then, he/she suggests going ahead with some further measures like taking a blood test or maybe consulting a specialist in that particular field. In scenario mentioned previously, there is time which was in a sense wasted. Speaking from personal experience the first situation involving doctor's prediction, he/she prescribes medicine which needs to be carried on for 4 days to maybe a week. If this guess is correct then viola! It works, but if it fails then the time spent on following that particular medicine course is completely wasted. To add to this there are even side effects a medicine could give you.

In India alone there are 13.9 Lakh cancer cases, out of which 64+% of them were detected in the later stages. For a disease like cancer, finding out about it in the earlier stage is considered super lucky that is the probability of it being cured is higher in the baby stages. Dr. Girdhar J. Gyani, Director General, anssociation of Healthcare Providers (India) spoke about major misdiagnosis in the field of medicine in the country. A Harvard study by Prof Jha proves that about 5.2 million (52 Lakh) medical errors are happening in India annually. A British journal also quoted that India, unlike other rapidly developing countries has maintained a lot of medical errors. This makes one question if this is a thing just in India or any developing country as there aren't many trained doctors and nurses to provide maximum hospital or clinic outcomes?

To negate the point, The National Academies of Sciences, Engineering, and Medicine estimated that around 12 million Americans receive misdiagnoses on a year basis, to which A BBC article added notes saying that diagnostic errors have caused about 40,000 to 80,000 deaths annually. Earlier in October this year a seven-year-old girl from Texas died due to major misprediction of her sickness by the paediatrics. UQ's Professor Ian Scott from Australia said that statistics have revealed around 21,000 cases of misdiagnosis made in a clinical setting involved serious harm to the life of the patient, and up to 4000 resulted in death. From looking at all these case studies, one out of every seven diagnoses is wrong whereas one out of every ten could even cause a patient his/her life.

3. DATA SCIENCE IN HEALTHCARE

Data Science is a field of study or research which uses scientific methods and algorithm to classify, predict or just give informative insights about a collection of records known as data which can be well-structured or

unstructured. Big data has now become an important tool for almost all the businesses and companies of any size. Due to many such reasons Data Science has proven to be the revolution by having an application of it, in any field you can ever remember.

One major application of this modern-day tool data science is in the field of Health and Medicine. From reading MRI scans or X-ray to Genomics, healthcare is considered to be one of the most effective and efficient use of Data Science and Machine Learning. The Medicine and Healthcare industry has heavily used the field of data science to improve the style of living by predicting any illness at a comparably early stage and maintaining a simple but at the same time complex health record of tons of patients which was never possible in an all human working system.

Predicting the disease of a patient just by look the past medical history and current symptoms shown by him/her using the concept of Machine Learning and Data mining is an ongoing struggle hopping for nothing but change or rather progress in it. Many techniques or Machine Learning algorithms have been developed to develop a model which would solve this problem. Though it works pretty decent till a certain end but not so well that it could be implemented fully in real life situations. Models such as Support Vector Machine (SVM), K-nearest neighbor (KNN), Decision Tree (DT), Logistic Regression (LR), AdaBoost (AB), Naïve Bayes (NB), fuzzy logic (FZ) have been developed and are broadly used in heart disease diagnostics. It is due to these Machine Learning Models which follow a medical expert like decision making there is a noticeably decrease in deaths caused by heart diseases alone.

4. USING MODELS

4.1 Logistic Regression:

Logistic Regression is one of Algorithms where analysis is conducted in regard of the dependent variable. [1] National Library of Medicine defines Logistic Regression models as "statistical models which describe the relationship between a qualitative dependent variable (that is, one which can take only certain discrete values, such as the presence or absence of a disease) and an independent variable."[2]. Similar to other regression models, this gives a probability of input variable lying in one of the classes. Talking about probability, the range of output lies between 0 to 1. [3] Types of Questions Binary Logistic Regression can answer:

1) What is probability of X female to like diagnose breast cancer (yes/no) of 'y' age.
2) Based on given features, can a person be classified as Middle Class or Lower class
3) Studying past medical history, what is probability of a person to catch some disease? And many more to use LR model as a binary classifier there comes a parameter called Threshold value, which will act as a boundary to predict classes. For e.g. for a binary classifier threshold value is 0.5 then any value higher than 0.5 will be considered in Class a Else Class B. A research survey conducted by Lavie et al, they gathered data from 2677 adults referred with suspected sleep apnoea, which they co-related it with absence or presence of hypertension. They considered 4 major features to introspect the odds ratio of person being associated by hypertension.

Analyzing the results of Odds ratio, 95% of Confidence interval is from 0.97 to 1.47 and hence we can say that age

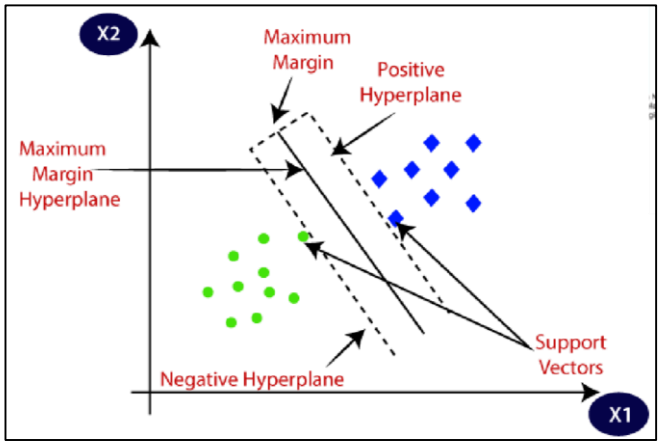

Fig. 1: Support Vector Machines

We interpret the age coefficient by saying that if we had doesn't much contribute in predicting hypertension the features which we considered: Age, Sex(Male/Female), BMI, Apnoea Index

Table 1: Risk factors for Hypertension

Risk Factor	Estimate (log odds)	(95 % CI)	Odds Ratio
Age (10 years)	0.805	0.718 to 0.892	2.24
Sex (Male) 0.161	-0.161 to 0.383	1.17	2.09
BMI (5 kg/m²)	0.256 to 0.409	0.332	1.39
Apnoea Index (10 units)	0.116	0.075 to 0.156	1.12

Two people of the same sex, and given that their BMI and apnoea index were also the same but one subject was 10 years older than the other, then we would predict that the older subject would be 2.24 times more likely to have hypertension

4.2 Support Vector Machine (SVM)

SVM's classify both linear and non-linear data. SVM algorithm does mapping and plotting of data points in ndimensional space where n is the number of features in a

dataset. SVM the finds a hyperplane which divides the points into 2 distinct classes by point of maximizing the distance from margin to support vectors.

For e.g., if two variables of dataset are being plotted in 2dimensional space then the separating hyperplane would be of 1 dimension, diving the space into half.

SVMs are highly used in Medical application fields wherein diagnosing a diabetic patient was a purpose. A study conducted by BMC medical informatics and decision making [5] claims that SVM is most promising classification approach for detecting persons for common diseases such as diabetes. The objective was to classify (diagnosed or undiagnosed diabetes vs. no diabetes or pre-diabetes. The data used for this study was from National Health and Nutrition Examination Survey (NHANES). The people under study were asked several sets of questions, "Have you ever been told by a doctor or health professionals that you have diabetes?", for model training, 14 features (associated with risk for diabetes) such as age, gender, physical activity, smoking, income, education, etc., were considered.

This data was being trained into SVM model and accordingly changes were constantly being done such as normalizing parameters, encoding data and to name a few. Test data were being used to assess the performance of models. On the basis of Sensitivity, specificity, PPV, NPV; the model yielded best performance by using RBF kernel function for 8 variables--family history, age, race and ethnicity, weight, height, waist circumference, BMI, and hypertension. The overall AUC value was 83.47%

4.3 Decision Tree:

DT is one of the most prominent and widely used algorithm in all sectors for all purposes. Decision Tree starts from root node till the leaf nodes, expanding on the basis of classifying features/conditions. Normally Decision Trees have multi-level nodes, wherein top most is root node and other internal nodes are test data where according to the outcome of node, further child nodes are being expanded until it reaches leaf node. A study conducted by some group of individuals; they developed a decision tree for detecting dengue. The data was being gathered by monitoring 1200 patients [6] who were having acute febrile illness and followed up for 4-week period. Over the analysis, 364 were dengue RT-PCR positive; 173 had dengue fever, 171 had dengue hemorrhagic fever, and 20 had dengue shock syndrome as final diagnosis. By using this data, a C4.5 decision tree was built by taking into considerations of all clinical, hematological and virological data. It was then tested by different data who accuracy turned out to be 84.7%.

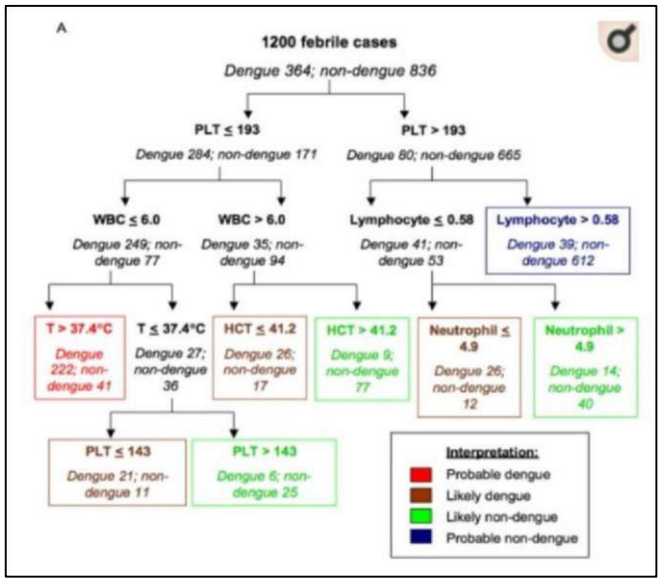

Fig. 4.3 Decision Tree

As we can observe in **Figure 4.3,** A Decision algorithm for predicting dengue diagnosis calculated on 1200 patients with data obtained in the first 72 hours of illness.

PLT=platelet count; WBC=white blood cell count; T=body temperature; HCT=hematocrit; Lymphocyte=absolute number of lymphocytes; Neutrophil=absolute number of neutrophils. The prediction of the algorithm is shown in colors: Red indicates probable dengue; brown indicate likely dengue; green indicates likely non-dengue and blue indicates probably non-dengue. B. Statistical (chi-square) analysis of splitting criteria performed on each subgroup at the decision nodes. OR=odds ratio; CI=95% confidence interval.

5. BENEFITS OF DATA SCIENCE IN HEALTHCARE

There are many advantages of Data Science in the field of Healthcare:

1) **Early prediction:**

Any diseases can be cured if they are detected in early stage. As it is always said "Prevention is better than cure." Knowing what illness is faced by a patient in the initial days will always help in the cure of the same. This will not only give the doctors the time but also increase the probability of a patient to be cured thereby saving life and improving general human lifestyle.

2) **Well Trained Models:**
The model will be trained on a huge amount of data, i.e. data of the whole nation or maybe the entire world, whereas the common doctor is trained only on his/her patients. In the doctor's case, each doctor has a very different period of experience and the experience (in case of the doctor) or amount of data it is trained on (in case of the algorithm) is an important attribute when it comes to the prediction of the sickness. Whatever be the scenario the model will always have a better efficiency when compared to the existing system.
3) **Improvement Diagnostic accuracy and efficiency:** Any machine learning model gets accurate over the period of time because of it is trained well on a data and then tested. Since this model will be trained more than any doctor has ever been as there are so many pre-recorded cases in the country or internationally, accuracy will be always greater. This accuracy will result in overall efficiency by reducing misdiagnosis and thereby saving many lives.
4) **Right Treatment:**
Using this model, the right treatment to the particular disease which faced by the patient, will be provided. Major part of the medical sector is dependent on English Medicines. It has been noticed that consuming a lot of them can lead to kidney failure. Some of these medicines have side effects too. When compared to the existing system, if a doctor's predication is wrong after consuming some medicines for a half a week or two another set of medicine are prescribed to the patient. Note that the previously prescribed medicines will now be considered as useless. The chances of this happening in case of the model are extremely less.
5) **Learning from mistakes:**
A Machine Learning model learns from its mistakes, this model would probably do some wrong predictions in early stage as the experience or training received by it is less but then it would learn from these mistakes and give improved results next time. This will improve its accuracy and thereby give a major rise to the efficiency factor of this model.

6. CHALLENGES FACED

However, the model currently faces a lot of challenges or has some drawbacks which are mentioned below:

1) **Huge Data:**
For countries like India where the population is huge or while considering the entire world, handling or rather maintaining data would be really difficult. This would be faced by the data scientists or ML engineers. To add to this, a human body daily produces a data equivalent to two terabytes. Storing this amount of data and handling it could be a tedious task.
2) **Difficulty in data cleaning and preprocessing:** As the size of data is huge, data cleaning won't be really easy and so is the process of data preprocessing. There is huge chance of the existing data to have missing values for factors like current health trend i.e. is it a viral season or the era is under a pandemic, etc. It will be really difficult for the ML engineers and data scientists to deal with it and thus giving rise to the complexity of developing this.
3) **Invasion of Privacy:**
There are certain laws in medical records to maintain the privacy of a patient but then when it comes to big data sharing for Data Science to exist in this field there are no rules to be followed. Sometimes a patient would like to keep his/her medical history private or hidden this could be a drawback of this system.

This change implemented on a larger scale could end up replacing the job of physicians or doctors in general which would definitely affect the economy.

7. CONCLUSIONS

The ultimate goal is to facilitate coordinated and wellinformed health care systems capable of ensuring maximum patient satisfaction. In developing nations, predictive analytics are subsequent big idea in medicine –the next evolution in statistics – and roles will change as a result. Patients can get to become higher knowing and can get to assume a lot of responsibility for his or her own care, if they are to make use of the information derived.

Physician roles can probably modification to a lot of an advisor than head, who will advise, warn and help individual patients. Physicians might notice a lot of joy in apply as positive outcomes increase and negative outcomes decrease. Perhaps time with individual patients can increase and physicians will another time have the time to create positive and lasting relationships with their patients. Time to assume, to interact, and to really help people; relationship formation is one of the reasons physicians say they went into medicine, and when these diminish, so does their satisfaction with their profession. Hospitals, pharmaceutical corporations and insurance suppliers can see changes furthermore. These changes which will virtually revolutionize the manner drugs are practiced for higher health and un-wellness reduction.

8. FUTURE SCOPE

1) Inculcation of automation in aspects of regular functioning too can be deployed as well.

2) User experience can be increased by enhancing personalized content on each user's dashboard or notification tags.
3) We will try to convert our model into a web application with many additional features.
4) A forum where User can discuss the issues with doctor and will also fix the appointment with the doctor based on available dates.

REFERENCES

[1] https://www.statisticssolutions.com/whatislogistic-regression/
[2] https://pubmed.ncbi.nlm.nih.gov/18450055/#:~:text=The%20Medical%20Subject%20Headings%20(MeSH,presence%20or%20absence%20of%20a
[3] https://www.statisticssolutions.com/whatislogistic-regression/
[4] https://www.healthknowledge.org.uk/elearning/statistical-methods/specialists/logisticregression
[5] https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-10-20
[6] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2263124/