GLIMPSE - Journal of Computer Science • Vol. 1(2), JULY-DECEMBER 2022, pp.27-31

A MACHINE LEARNING BASESD HEALTHCARE DIAGNOSTIC MODEL

Anurag Gupta
Assistant Professor, Ajay Kumar Garg Engineering College, Ghaziabad, U.P India
guptaanurag@akgec.ac.in

***Abstract*— The current pandemic has taught people to focus more on their health and continuous monitoring of their health has become need of the hour. So, people need to be more concerned about their health. Unfortunately, nowadays the doctor human resource is lesser than the patient. Using Information Technology can definitely prove to be a boon in healthcare sector. Information Technology will produce more medical data, which will give birth to multiple fields of research. Many efforts are done to cope with the explosion of medical data on one hand, and to obtain useful knowledge from it on the other hand. Techniques like predictive analytics and machine learning will help to extract useful knowledge and help in making decisions. With the use of machine learning algorithms, our model is able to predict the disease and is also able to anticipate its cure. This paper introduces a chatbot for collecting the information and later providing the patient with vital information regarding their disease and cure. This paper discusses the potential of utilizing machine learning technologies in healthcare and outlines various industry initiatives using machine learning initiatives in the healthcare sector. This paper also focuses on saving the data of patient, and also providing him with a suitable specialist doctor for his health issues.**

Keywords- Machine learning, healthcare, Prediction, Chatbot.

I. INTRODUCTION

When epidemiologists work to mitigate the effects of a pandemic, they study the patients' populations and identify causes, risk factors, suitable treatments and finding the variation in the disease. Procedures of randomized controlled trials and case studies become the basis for evidence-based medicine. However, such methods are time taking and involve a lot of money, might be biased and not fulfill the desired results, and hence the outcomes may not be applicable to real-world patient populations. Such studies are not easy to execute over big patient population.

Internationally, the countries are adopting to access electronic health records (EHRs) a lot due to strategies and agencies that incentivize their use such as the Health Information for Economic and Clinical Health Act in the United States, the National Agency for Health IT in Denmark, and the National e-Health Authority in India.

Data analysis approaches are of many types like descriptive, exploratory, inferential, predictive and casual. A descriptive evaluation depicts a summarization of information without interpretation and an exploratory evaluation looks for interconnections between variables in a dataset. An inferential observe measures the degree to which an observed association in a populace will stand different from the dataset from which it became derived, and a predictive analysis attempts to the opportunity of the final results at the extent of an individual. Finally, a causal analysis determines how modifications in a single variable affect another.

It is quite difficult to identify the type of question being asked in a given study to determine the type of data analysis that is appropriate to use in answering the question.

So, in order to do the above tasks we have used Decision Tree as a methodology to predict the disease of the patient based on symptoms and patient history, the cure and also a doctor based on what specialist he requires. Decision tree being a supervised methodology will use label from the dataset in order to train it and thereby producing accurate results. The most important feature of the model is a chatbot, or the "health assistant" that will eventually help in collecting the data first and after decision tree performs its function at the backend, it gives the results of the ML technique to the user.

Hence, the system provides text assistance you can communicate with bot like user friendly. Further The Bot provides which type of disease you have based on user symptoms and appeared doctor details respective to user disease. The chatbot will also clarify the users' symptoms by collecting several relevant questions and the symptom confirmation will be done. Based on this the disease will be categorized as minor and major disease. Chatbot will reply whether it is a major or minor disease if it is a major disease user will be suggested with the doctor details for further treatment.

II. LITERATURE SURVEY

The author in [1]proposed this methodology that is usedin support vector machine algorithms, NLP, Word order similarity among sentences. The Heart - disease dataset is used. The methodology is used because SVM can solve more complex problem classification and further training is easier. Its disadvantage is NLP incorporation.

The author in [2] describes a model using Ensemble learningin which the Classifier is trained in a single iteration eachand final choice is made by the major vote. Stacking-based ensemble learning is also used where the major voting ensemble acts as combiner. General health dataset and the Pima Indian diabetes dataset are used. Advantage of this model is that no classifiers is dominating among the various classifiers and its drawback is that the computation and design time are high. The author in [3] used Knowledge graph and hierarchical bi-directional attention methodology in which architecture of hybrid QA model Framework, combines a knowledge graph to manage a medical dataset and HBAM to understand the text. Dataset of 3,500 entities (which include 675 diseases and 2825 symptoms) and 4,500 relationships is used. Its advantage is that it makes use of structured memory so that it may help to make the task maintenance and extractionof Domain-specific knowledge easier and its disadvantage is that the work is complex here. The author in [5] supports NLU, NLTP, Multinomial Naive Bayes methodologies are used in which Sentiment Analysis, Tokenization, Named Entity Recognition, Normalization, Dependency Parsing are used. Corpus_words dataset and class_words dataset is used. Its advantage is that its simple to build and its disadvantage is that no data provided for the disease. In [6] RNN , NLP , speech to text methods are used in this model . Implementation of Sequence-to- Sequence Model ,Apriori algorithm are being done .This model is trained on dataset available from the New York Presbyterian Hospital. This apriori principle can reduce the number of items we need to evaluate. This model requires a lot of time for training even though the hardware is capable of handling it. This model [7] used Teacher forcing methodin their proposed system. A medical advice generator and a general empathetic conversation generator with four parallel LSTM layers are used bin this paper. Also, Concatenation, Facebook AI Empathetic Dialogue dataset and Medical Question answering dataset is used. Accuracy of intent classifier was 98.5% and that of emotion classifier was 92.4%. Poormodel and performance instability. The paper [8] proposes a model that implements Text mining with AI and uses API Medic methods. GloVe vectors and APIMedic algorithms are used in this. A survey of demographic information , a natural language description of symptoms, further elaboration on the symptoms and the presumed diagnosis and ApiMedic database. It completes a database from API medic and easier to check symptoms . It does not provide accurate results. [11]It features conversational knowledge based on knowledge-graph for factoid medical questions, using algorithms like natural language interpreter, dialog manager, and natural language generator. RDF dataset is used in it. It efficiently handles the dialog, marks missing information, and generates more precise and contextualized responses. [13] This model has a Facebook chatbot for sexual health information on HIV/AIDS. It uses NPC Editor to collect chatbot responses, a dialog manager, and plug-in to Facebook. Algorithms performing classification and NLP algorithms are helpful for this task. It uses online survey, QA in SHIHbot Domain as a dataset. The live conversations will depict SHIH boat's ability to understand new questions, the chatbot's ability to tackle with the marked questions outside of the domain knowledge, and the overall flow of dialog. A Survey [14] on chatbot implementation in health care using NLTK and algorithm like Natural language processing is done. Dataset is a QA record in it. It is user friendly and can be used by any person who knows how to type in their own language in mobile app or desktop version, and provides personalized diagnosis based on symptoms.

III. MODEL ASSUMPTIONS

The following section discusses the major libraries and dataset used in implementing the model. Chatbots are a new age assistants for people such that demands can be met at the right time, and wait time of people for actual human assistants can be reduced. A healthcare based model consisting of such technique can help the overwhelmed healthcare systems to operate in a better way

A. Platform Setup

In this paper Python language is used. Python is an interpreted high-level programming language for the general-purpose programming. The syntax of this language is very simple as compared to English language. The use of python is diverse in the field of software to create workflows, can be used to connect to the database, to handle big data and perform complex mathematics. For making the GUI as well, python language is used.

B. Model Libraries

- Numpy- It is the basic package for scientific computing. It used as an efficient multi-dimensional container of collective or comprehensive data.
- Pandas- open source library providing high- performance, easy-to-use data structures and data analysis tools.
- Matplotlib- Python 2D plotting library that generates publication quality figures in a variety of hardcopy formats and interactive environments across platforms.
- TextBlob- Python library for processing textual data. It provides a simple API for diving into NLP tasks.
- Tkinter- It is a basic GUI library for Python. Python when combined with Tkinter provides a fast and convenient way to design

GUI applications. Tkinter gives a powerful object- oriented interface to the Tk GUI toolkit.

C. Dataset Generation

The sample dataset which we used to predict the diseases as per the symptoms selected by the user, as of now the dataset includes 4921 rows and 133 columns. This dataset is collected from Kaggle [15]. This includes 41 different types of diseases and their several symptoms. Also, we have a Doctor"s dataset [16] which will help in recommending a suitable doctor to the user according to his disease. He can visit the Doctor"s site and also book an appointment whenever he wishes to consult the doctor. Table 1 consists of the data description of data in Doctor Dataset. However the symptom dataset has 133 columns in Boolean form for all the symptoms and the disease.

TABLE I. DATA DESCRIPTION OF DOCTOR"S DATASET

S.no.	Data Colums	Data type
1.	Doctor's Name	String
2.	Website Name	Stirng

IV. IMPLEMENTATION AND INTERPRETATION

This section focuses on the basic architecture and work flow of the model proposed for healthcare diagnostics purpose.

A. Basic Framework

- As already mentioned this paper deals with healthcare based data. The first step here would be data cleaning and thereafter data training for the machine learning process using Decision Tree Classifier.
- Then, the chatbot is created whose response will be based on the classifier trained previously.
- After successful prediction of the disease the chatbot also provide the patient with a recommendation of specialist doctor for him or her.
- For visualization of the whole model, a GUI is created that will help the patient to use the model with ease.

B. Execution Process

- **Data Training-** The Project involves two different types of datasets namely the training dataset and the doctor dataset of which the first one is the training dataset consist of 133 columns or we can say 133 different types of symptoms on the basis of which we are going to find out the resulting prognosis. The prognosis are the class labels and in all there are 41 different types of diseases as of now. So based on the given symptoms we will conclude with a specific prognosis. This training data contains the data of 4921 different types of individuals which are suffering from different types of symptoms. These symptoms are labelled with Boolean values which means either the person is having that symptom or doesn"t have it. So here 1 means the person is having that symptom and 0 means the person doesn"thave that specified symptom. These all information are related to the training dataset. Coming to another dataset which is the doctor dataset. The Doctor dataset consists of 2 different columns of which the first column consists of the Doctor's name and the second column contains a respective website link from where the patients can book their appointment. This dataset contains 41 entries of doctors as of now. All these links have their origin from Practo.com which is a healthcare information technology provider based in Bengaluru, India. It is an online doctor consultation website/app that offers complete telemedicine solutions.Talking about training of our decision tree classifier model we splitted our primary data into 75% training and 25% testing dataset using train_test_split() method which is present in scikit-learn. This module is a freesoftware machine learning library for the Pythonprogramming language. It consists of various algorithms like regression, clustering and classification algorithms including random forests, supportvector machines, k-means, gradient boosting and DBSCAN, and is created to correlated with the Python numerical and scientific libraries which are known as NumPyand SciPy.

- **Dimensionality Reduction-** The input variables or features of a dataset is referred to as its dimensionality. Dimensionality reduction refers to the technique in which a higher dimensional data space can be transformed onto a lower dimensional space or we can say it reduces the number of input variables in a dataset. Dataset having more input features often make modelling tasks more challenging to model, so more generally referred to as the curse of dimensionality. The performance of machine learning algorithms can degrade with too many input variables. If your data is represented by rows and columns, such as in a spreadsheet then the input variables are columns that are fed as input to the model to predict the target variable. So, we can consider the column representing dimensions or an n-dimensional feature space. Hence it is often useful to reduce the dimensionality by projecting the data to a lower dimensional subspace which captures the essence of the data. There are many techniques that can be used for dimensionality reduction which are as follows: Feature Selection Methods, Matrix Factorization, Manifold Learning, Autoencoder Methods etc.

- **Classification-** Classification is a two-step process, learning and prediction. In the learning step, the model is developed based on given training data and in the prediction step, the model is used to predict the response from the given data. Decision Tree is one of the most popular and easiest classification algorithms to understand and interpret. This algorithm belongs to the family of

supervised learning algorithms which is used for solving regression and classification problems too. The goal of this classifier is to create a training model that can be used to predict the class or value of a target variable by learning simple decision rules derived from the training data. There are also some basic assumptions which are needed to consider while creating a decision tree such as in the beginning the whole training set is consideredas the root after that feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model. Records are distributed recursively on the basis of attributes values. Even placing attributes as root or internal nodes of the tree is done by some statistical approach. Decision trees are easy to interpret and visualize and it can easily capture Nonlinear patterns. It requires fewer data pre-processing from users hence there is no need to normalize columns. They are also used for feature engineering such as predicting missing values which are suitable for variable selection. The main disadvantage of the decision tree is that it is sensitive to noisy data as it can overfit noisy data so it is recommended that to balance out the dataset before creating the decision tree.

- **Training Bot with classification results-** The training bot is a python-based menu driven application which interacts with the user and asks the symptoms one by one. On the basis of the response which will be either „yes‟or „no‟ the training bot will parse the tree structure. Ifthe user responds „no‟ for a particular symptom the botwill recursively come with some other sets of symptomsand so on. But if the user answer „yes‟ for a particular disease then the bot will check all the related disease for better understanding of the prognosis and also parse the doctor‟s dataset for the Name of the doctor who is an expert in this field and provide the user with all the details including the website link using which the user can book an appointment with the doctor.

- **GUI Creation- Python** has a lot of GUIframeworks, but Tkinter is the only framework that‟s built into the Python standard library. It is cross-platform so the same code works on Windows, macOS and Linux. Tkinter is light weighted and relatively painless to use compared to other frameworks. There are various widgets like button, check-button, canvas, entry, etc. that are used to build the python GUI applications. The GUI starts with a window frame asking for either login or signup. Based on the choice selected, the respective window gets opened. The Sign in window consists of two text fields one for username and other is for password. If the user information is authentic then the system will allow the user to log into the system else the user will need to first register or create an account and after then he will be able to login into the system. In the same way the register window also consists of two text fields that allow the user to create his username and password. Once the user logs into the system after proper authentication, the symptoms‟ window pops up for user to insert the symptoms into the fields. Thereafter, when the model is fed with an appropriate number of inputs, it then generates a response in the form of the predicted disease, symptoms given, confidence interval and the recommendation for the doctor to visit next. The symptoms window also provides a link to book an appointment with the concerned doctor which can be copied and pasted into the browser by the user for further operations. Fig1 and Fig 2 describe the execution processes in diagrammatically.

V. EXPERIMENTAL OUTCOMES

A. GUI Implementation

Figure 3 shows the GUI developed as a result of the proposed methodology. The GUI will help the user to enter inside the system. The user will login in case of an already registered user. The user shall register if he/she is visiting the application for the first time. Hence, on the basis of diagnosis, the chatbot replies on the GUI. The GUI has been created using tkinter module in python. GUI in python is not heavy and takes less time to process. The GUI‟s response time was also better.

B. Results-

The chatbot will respond in a similar way as shown in Fig 4, the GUI is already discussed in the previous section and Figure 3. The training accuracy of the decision tree was found to be 98% and the testing accuracy was 96%. Hence, we were able to effectively classify and predict diseases based on their symptoms. The high accuracy would help in efficiently managing the patients and resolving their queries by an Artificial Intelligence method.

VI. CONCLUSION

The proposed methodology will automate the process of diagnosing a disease and will also help the overwhelmed medical infrastructure in times of crisis like a pandemic. The Decision tree algorithm used for classification plays an important role in predicting the disease and also later the recommendation module of doctors to the patients make it easier for the patient to find the perfect solution to his/her disease. Automation in the medical field is the need of the hour. Chatbots and other AI tools shall help in creating a mechanism such that mild and moderate symptoms of a disease could be treated at home and leaving the hospitals for serious patients.

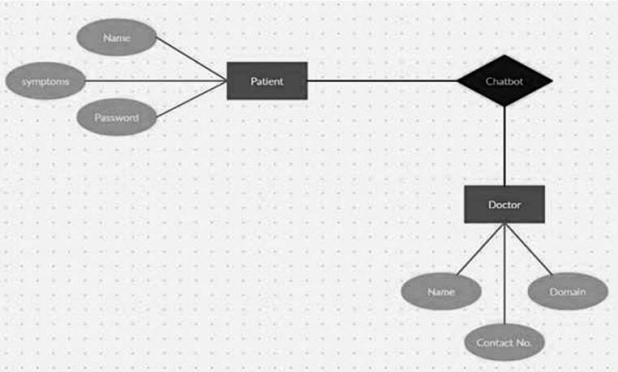

Fig. 1. Data associated with the chatbot.

This paper shall be helpful in such automation and hence further modification of the method can be more useful and efficient. This paper focuses on both the methodology and GUI, such that it becomes easy for the user to get his disease diagnosed at the earliest and also, connecting to the most suitable doctor also becomes easy.

```
In [1]: runfile('C:/Users/HP/Downloads/Health Care the diagnostic centre (1).py', wdir='C:/Users/HP/
Downloads')
Answer with Yes/yes or No/no for the symptoms
slurred_speech ?

no
pain_behind_the_eyes ?

yes
['You may have Dengue']

symptoms present  ['pain_behind_the_eyes']

symptoms given ['skin_rash', 'chills', 'joint_pain', 'vomiting', 'fatigue', 'high_fever', 'headache',
'nausea', 'loss_of_appetite', 'pain_behind_the_eyes', 'back_pain', 'malaise', 'muscle_pain',
'red_spots_over_body']

The model suggests:

Consult  ['Dr. Ajay Jain']

Visit  ['https://www.practo.com/delhi/doctor/dr-ajay-jain-ear-nose-throat-ent-specialist-1?
specialization=Ear-Nose-Throat%20(ENT)%20Specialist&practice_id=664069']
```

Fig.2. Chatbot Implementation

REFERENCES

[1] Dharwadkar, Rashmi, and Neeta A. Deshpande. "A medical chatbot." Int J Comp Trends Technol 60.1 (2018).
Bali, Manish, et al. "Diabot: a predictive medical chatbot using ensemble learning." Int. J. of Recent Technol. and Eng. (2019): 2277- 3878.

[2] Bao, Q., Ni, L. and Liu, J., 2020, February. HHH: an online medical chatbot system based on knowledge graph and hierarchical bi-directional attention. In Proceedings of the Australasian Computer Science Week Multiconference (pp. 1-10).

[3] Kalla, D. and Samiuddin, V., 2020. Chatbot for medical treatment using NLTK Lib. IOSR J. Comput. Eng, 22.

[4] Chung, K. and Park, R.C., 2019. Chatbot-based heathcare service with a knowledge base for cloud computing. Cluster Computing, 22(1), pp.1925-1937.

[5] Harilal, N., Shah, R., Sharma, S. and Bhutani, V., 2020. CARO: an empathetic health conversational chatbot for people with major depression. In Proceedings of the 7th ACM IKDD CoDS and 25th COMAD (pp. 349-350).

[6] Aswini, D., 2019. Clinical medical knowledge extraction using crowdsourcing techniques. Int. Res. J. Eng. Technol, 6.

[7] Rarhi, K., Bhattacharya, A., Mishra, A. and Mandal, K., 2017. Automated medical chatbot.

[8] Zarouali, B., Van den Broeck, E., Walrave, M. and Poels, K., 2018. Predicting consumer responses to a chatbot on Facebook. Cyberpsychology, Behavior, and Social Networking, 21(8), pp.491-497.

[9] Dahiya, M., 2017. A tool of conversation: Chatbot. International Journal of Computer Sciences and Engineering, 5(5), pp.158-161.

[10] https://www.kaggle.com/data/86712

[11] https://www.kaggle.com/pawanyalla/medical

[12] Yoo, S. and Jeong, O., 2020. An Intell Model and Knowledge Graph. Journa Studies, 24(3).

[13] Bohle, S., 2018. "Plutchik": artificial intelligence chatbot for searching NCBI databases. Journal of the Medical Library Association: JMLA, 106(4), p.501.

[14] Brixey, J., Hoegen, R., Lan, W., Rusow, J., Singla, K., Yin, X., Artstein,

[15] R. and Leuski, A., 2017, August. Shihbot: A facebookchatbot for sexual health information on hiv/aids. In Proceedings of the 18th annual SIGdial meeting on discourse and dialogue (pp. 370-373).

[16] Sophia, J.J., Kumar, D.A., Arutselvan, M. and Ram, S.B., 2020. A survey on chatbot implementation in health care using NLTK. Int. J. Comput. Sci. Mob. Comput, 9.

ABOUT THE AUTHOR

Mr Anurag Gupta is working as an Assistant Professor in CSE Department in Ajay Kumar Garg Engineering College, Ghaziabad. His area of interest is Machine Learning, AI and Deep Learning