Disease Prediction System

SHUBH PATEL[1], ASHUTOSH YADAV[2], NISHANT UPADHYAY[3]
[1, 2, 3] *Information Technology Department, Thakur college of engineering & technology, Mumbai, India*

Abstract— Predire is a web-based application that will help to predict diseases and provide a educational environment for students and patients. This project will provide an effective platform for the medical students and patients. There's been an increase in soul disease prediction models but there has never been models which predicts more than one variety of disease in a wider spectrum. This project focuses and tackles this research problem by using classification algorithm of data science on meticulously curated medical dataset. Random Forest Classifier is used for the classification of the dataset, where the users predict their disease based on 5 input symptoms. The met conclusion of this project is to predict user's diseases accurate, however the research is labelled to be in progress as with more medical data, more accurate, descriptive and intricate prediction models can me made.

Index Terms- Healthcare Prediction, Recommendation, Machine Learning.

I. INTRODUCTION

Predire is a website that uses Machine Learning algorithms to predict diseases with the help of symptoms provided by our users/patients. It also provides Healthcare related news and research papers for doctors and medical students.

This website caters not only to the public but also to hospitals and private corporations.

With the integration of powerful machine learning techniques and accurate medical data, we have developed a website that can predict 41 diseases. It predicts diseases that range from common diseases such as common cold, fever, etc to severe diseases such as hepatitis, tuberculosis, diabetes, etc.

The already existing applications for disease prediction do not predict a wide variety of diseases, instead, they are focused on predicting a single disease. Additionally, they do not provide an environment where the user can learn or gain knowledge about topics related to their application.

Predire tackles all the above constraints and provides an intricate and accurate prediction using a different approach than the pre-existing systems on the internet. Predire's prediction model is weight-based. It predicts the highest weighted disease based on 5 symptoms provided by the user. Here, weights are values ranging from 1-5 assigned to each disease based on their severity, where the higher value indicates a higher severity.

II. THEORY

The main objective of making this project (Predire) is to create an environment where the user can browse through various Health-related news, site curated research papers and get an accurate synopsis of the disease they might have while using our Prediction page. As this is a Web Application, it is easily accessible from anywhere in this world as long as there is a stable internet and hardware, and software requirements are met. The User Interface of the website is meticulously designed for a user-friendly experience wherein the user can easily navigate through the website using our well-designed Home page and Navigation bar. The main highlight of the website is our Prediction model. It is an easy-to-understand model which consists of 5 symptom inputs and a predict button. We have also provided detailed instructions on how to use the model for the user's convenience. There is also a Blog/News section where we provide news feeds from around the world related to the field of medicine. This project is also trying to create an environment that caters to medical students by providing latest research papers in the field of medicine.

III. LITERATURE SURVEY

We read a variety of research articles as our project was being developed, and they helped us comprehend our alternatives and potential solutions. They also assisted us in comprehending the technical and mathematical knowledge that will be useful.

One of the papers that we studied was the "Disease Prediction in Data Mining Technique" by the Authors S. Vijyaranhi and S. Sudha and it talked about how data mining techniques are used to predict various types of diseases. This paper reviewed research papers that focused on predicting heart disease, diabetes, and breast cancer. The prediction of heart disease was discussed using machine learning algorithms such as naive bayes, K-NN, and Decision List. When compared to other algorithms, the nave bayes algorithm has the highest classification accuracy. The author concluded that naive bayes correctly classifies 74% of the input instances. Following that, we will talk about breast cancer prediction. It is carried out using a variety of data mining techniques, including C4.5, ANN, and fuzzy decision trees. Using C4.5, the author discussed and resolved the problem's issues and algorithms. Using ANN, the author concluded that the network is trained to have consistent accuracy over time and good performance. Finally, we discuss diabetes prediction, where the author discovers overfitting and over generalization behavior of classification using a homogeneity-based algorithm. The author predicts class accuracy using a genetic algorithm.

The second article was titled "The Use and Role of Predictive Systems in Disease Management" published by the authors David H. Gent,, Walter F. Mahaffee, Neil McRoberts, and William F. Pfenderdid a study about the Disease predictive systems are intended to be management aids. With a few exceptions, these systems typically do not have direct sustained use by growers. Rather, their impact is mostly pedagogic and indirect, improving recommendations from farm advisers and shaping management concepts. The degree to which a system is consulted depends on the amount of perceived new, actionable information that is consistent with the objectives of the user. Often this involves avoiding risks associated with costly disease outbreaks. Adoption is sensitive to the correspondence between the information a system delivers and the information needed to manage a particular pathosystem at an acceptable financial risk; details of the approach used to predict disease risk are less important. The continuing challenge for researchers is to construct tools relevant to farmers and their advisers that improve upon their current management skill..

And the third paper published by the authors Mangesh Limbitote and Kedar Damkondwar named "Prediction Techniques of Heart Disease using Machine Learning" talks about Heart is one of the most important part of the body. It helps to purify and circulate blood to all parts of the body. Most number of deaths in the world are due to Heart Diseases. Some symptoms like chest pain, faster heartbeat, discomfort in breathing are recorded. This data is analysed on regular basis. In this review, an overview of the heart disease and its current procedures is firstly introduced. Furthermore, an in-depth analysis of the most relevant machine learning techniques available on the literature for heart disease prediction is briefly elaborated. The discussed machine learning algorithms are Decision Tree, SVM, ANN, Naive Bayes, Random Forest, KNN. The algorithms are compared on the basis of features. We are working on the algorithm with best accuracy. This will help the doctors to assist the heart problem easily.

Furthermore, in the fourth paper named "Data-driven Automatic Treatment Regimen Development and Recommendation" by authors Leilei Sun, Chuanren Liu, Chonghui Guo, Hui Xiong, Yanming Xie discussed about the analysis of EMR records to detect typical treatment regiments and measuring (quantitatively) the effectiveness of those regimens for specific patient cohorts. The authors compare the similarity of treatment records in the EMR, use Map Reduce Enhanced Density Peaks based Clustering to group similar ones to treatment regimens, extract semantically meaningful information for the doctor, and estimate the treatment outcome for a patient cohort for a typical treatment regimen. The results of an empirical study using this approach show that the patient's effective rate and cure rate both increases.

In summary, recommendation systems are used to suggest various measures to patients or users based on their history or symptoms. Therefore, this research

helps us understand different aspects of customer behaviour.

IV. METHODOLOGY AND PROPOSED MODEL

When creating a project or piece of software, it is essential to comprehend the process or kind of software development lifecycle model that will be used. We need to select the SDLC model that will work best for our project from a variety of models, which is why we use the agile methodology.

Additionally, the actions taken up until this point have been previously discussed. After deciding on a project title, we first began studying about and establishing the project's requirements. During this, we also decided on the approaches to be applied. After that, we made the decision to research already-in-use systems to gain a deeper knowledge before implementing them. We devised a schedule for when this implementation should take place, along with considerations for any revisions that could be necessary and the errors that should be investigated. Therefore, we selected Agile methodology.

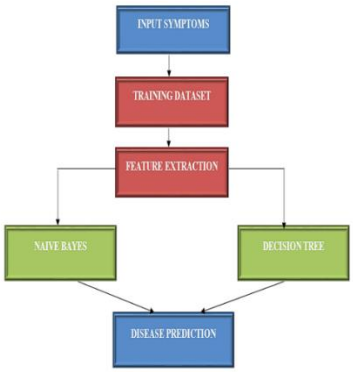

Fig. 4.1 Data flow diagram

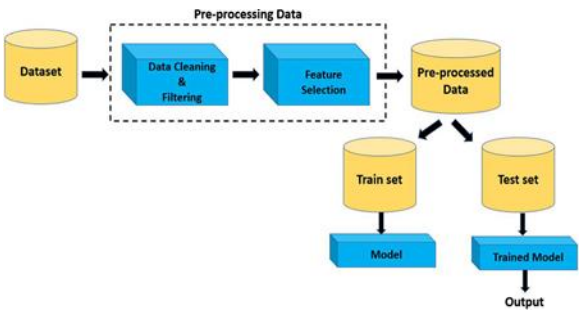

Fig. 4.2 Block diagram

Above attached diagrams are a couple of the ones that we made during our project planning phase using tools available online to achieve our goal.

V. IMPLEMENTATION

This system's primary goal is to diagnose illnesses using user-provided symptoms as input. Additionally, if it is feasible, we might test a few additional machine learning models so that we can compare how they perform to the ones we are currently using.

We decided on Logistic Regression algorithm along with feature selection to yield high accuracy using training and testing dataset which was freely available for us to use. The model was trained and tested using the dataset which contains the columns of different symptoms and based on these symptoms the prediction of diseases.

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt
        from sklearn.model_selection import train_test_split
        from sklearn.svm import SVC
        from sklearn.ensemble import RandomForestClassifier
        from sklearn.metrics import f1_score, accuracy_score, confusion_matrix
        import seaborn as sns
        from tkinter import *
        from tkinter import messagebox
        from tkinter import ttk
        import sys
        import urllib
        import urllib.request
        import pickle
        import os
```

Fig. 5.1 Libraries Used

```
In [2]: df = pd.read_csv('dataset.csv')
        df.head()
```

Out[2]:

	Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4	Symptom_5	Symptom_6	Symptom_7	Symptom_8	Symptom_9	Symptom_10
0	Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches	NaN	NaN	NaN	NaN	NaN	NaN
1	Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	Fungal infection	itching	skin_rash	dischromic _patches	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	Fungal infection	itching	skin_rash	nodal_skin_eruptions	NaN	NaN	NaN	NaN	NaN	NaN	NaN

```
In [6]: cols = df.columns
        data = df[cols].values.flatten()

        s = pd.Series(data)
        s = s.str.strip()
        s = s.values.reshape(df.shape)

        df = pd.DataFrame(s, columns=df.columns)
        df.head()
```

Fig. 5.2 Reading and training the dataset

Figures 5.1 and 5.2 are a few code examples that we used and put into practise when categorising the dataset during training and testing and when using feature engineering, respectively.

VI. RESULTS AND DISCUSSION

The model predicts the diseases and makes a confusion matrix and based on the predictions results are provided. Based on the results of confusion matrix using different algorithms results are displayed. Below is the output that we managed to achieve.

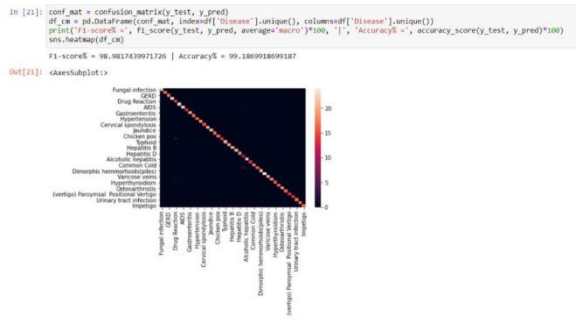

Fig. 6.1 Confusion Matrix

Fig. 6.2 Results

All the above-mentioned actions were taken after consulting the manual and receiving advice from professionals in the field. What is anticipated of the system is predicted by its output. One anomaly, which is that the values that are to be submitted for prediction right now are being given and taken in normalized form, is something we would like to fix in the following phase. So, that's one significant feature that we anticipate changing later in the ensuing period.

The Website will be based on prediction model where users can embed their symptoms and get the possible prediction based on the symptoms. The best possible accuracy level of prediction model. Nearby hospital will be suggested based on the predicted disease. Users will have all the medical data stored in one place with all the prescribed data uploading facilities. The simple Gui will help user in easy navigation. Many other side features such as recommendation of nutritional foods, tracking of one's data hence acting as a data storage as well, appointment booking system, chatting system and many more minor features which are aimed to healthify one's life

The scope of the project is clear to give a simple and attractive application to simplify the work as well as to reduce the efforts while doing it offline or we can say by doing it with old methods. In this application we are able to save database of all patients present on the site.

Prediction Module: As part of healthcare, a prediction model is necessary as it would help users to know if they are suffering from disease or not thereby also reducing the cost of visiting a doctor which costs a lot.

Throughout this research paper, logistic regression, naive Bayes, support vector machines, decision trees, random forests, XGBoost classifiers, CatBoost classifiers, AdaBoost classifiers, and extra-tree classifiers. Experimental results show that there are his two ensemble learning methods, Adaboost classifier and XGBoost classifier. It is very difficult to predict the actual customer society. With the upcoming concepts and frameworks of reinforcement learning and deep learning, machine learning is proving to be one of the most efficient ways to tackle problems such as churn prediction with more accuracy and precision in the future. increase.

CONCLUSION

The mechanism proposed aims at the continuous data and establishes the prediction model based on the regression analysis method, which is not only applicable to the analysis and prediction of the guidance data in the smart medical industry. In the future, new features can be added to improve the accuracy of the prediction model. For example, new disease data have an impact on the number of systems seeking medical treatment and the hospital. In addition, in the face of a larger amount of data, we can use the cloud architecture in this paper to carry out distributed computing. It has all necessary features for one's need. So, this project will help consumers for improving their health and know more about digital healthcare. In the near future, smart medical healthcare can be improved in some aspects; for example, this way can help patients and doctors identify the right information and deal with this information effectively. The first step of literature survey and research was completed till now for more knowledge of the domain.

REFERENCES

[1] Analyn N. Yumang, Ericson D. Dimaunahan, Paulo Alfonso Borja, Ericson De Castro and Carlito Pablo, "Android Based Blood Pressure Monitoring System Using Wrist Sphygmomanometer", *10th International Conference on Biomedical Engineering and Technology (ICBET 2020). Association for Computing Machinery*.

[2] S. A. Riffat, F. Harun and T. Hassan, "An Interactive Tele-Medicine System via Android Application", *2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*, pp. 148-152, 2020.

[3] S. A. Riffat, F. Harun and T. Hassan, "An Interactive Tele-Medicine System via Android Application", *2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*, pp. 148-152, 2020.

[4] R. Lee, K. Chen, C. Hsiao and C. Tseng, "A Mobile Care System with Alert Mechanism", *IEEE Transactions on Information Technology in Biomedicine*, vol. 11, no. 5, pp. 507-517, Sept 2007

[5] A. Jayed Islam, M. Mehedi Farhad, S. Shahriar Alam, S. Chakraborty, M. Mahmudul Hasan and M. Siddat Bin Nesar, "Design Development and Performance Analysis of a Low-Cost Health-Care Monitoring System ", *2018 International Conference on Innovations in Science Engineering and Technology (ICISET)*, pp. 401-406, 2018.