Multiple Disease Prediction System

Ankush Singh[1], Ashish Yadav[2], Saloni Shah[3], Prof. Renuka Nagpure[4]

[1]Ankush Singh, Dept. of Information Technology Engineering, Atharva College of Engineering
[2]Ashish Yadav, Dept. of Information Technology Engineering, Atharva College of Engineering
[3]Saloni Shah, Dept. of Information Technology Engineering, Atharva College of Engineering
[4]Prof.Renuka Nagpure, Dept. of Information Technology Engineering, Atharva College of Engineering
Maharashtra, India

---***---

Abstract - Machine learning and Artificial Intelligence are playing a huge role in today's world. From self-driving cars to medical fields, we can find them everywhere. The medical industry generates a huge amount of patient data which can be processed in a lot of ways. So, with the help of machine learning, we have created a Prediction System that can detect more than one disease at a time. Many of the existing systems can predict only one disease at a time and that too with lower accuracy. Lower accuracy can seriously put a patient's health in danger. We have considered three diseases for now that are Heart, Liver, and Diabetes and in the future, many more diseases can be added. The user has to enter various parameters of the disease and the system would display the output whether he/she has the disease or not. This project can help a lot of people as one can monitor the persons' condition and take the necessary precautions thus increasing the life expectancy.

Key Words: Diabetes, Heart, Liver, Knn, Random forest, XGBoost.

1.INTRODUCTION

In this digital world, data is an asset, and enormous data was generated in all the fields. Data in the healthcare industry consists of all the information related to patients. Here a general architecture has been proposed for predicting the disease in the healthcare industry. Many of the existing models are concentrating on one disease per analysis. Like one analysis for diabetes analysis, one for cancer analysis, one for skin diseases like that. There is no common system present that can analyze more than one disease at a time. Thus, we are concentrating on providing immediate and accurate disease predictions to the users about the symptoms they enter along with the disease predicted. So, we are proposing a system which used to predict multiple diseases by using Django. In this system, we are going to analyze Diabetes, Heart, and malaria disease analysis. Later many more diseases can be included. To implement multiple disease prediction systems we are going to use machine learning algorithms, and Django. Python pickling is used to save the behavior of the model. The importance of this system analysis is that while analyzing the diseases all the parameters which cause the disease is included so it is possible to detect the disease efficiently and more accurately. The final model's behavior will be saved as a python pickle file.

1.1 Description

A lot of analysis over existing systems in the health care industry considered only one disease at a time. For example, one system is used to analyse diabetes, another is used to analyse diabetes retinopathy, and another system is used to predict heart disease. Maximum systems focus on a particular disease. When an organization wants to analyse their patient's health reports then they have to deploy many models. The approach in the existing system is useful to analyse only particular diseases. In multiple diseases prediction system a user can analyse more than one disease on a single website. The user doesn't need to traverse different places in order to predict whether he/she has a particular disease or not. In multiple diseases prediction system, the user needs to select the name of the particular disease, enter its parameters and just click on submit. The corresponding machine learning model will be invoked and it would predict the output and display it on the screen.

1.2 Problem system

Many of the existing machine learning models for health care analysis are concentrating on one disease per analysis. For example first is for liver analysis, one for cancer analysis, one for lung diseases like that. If a user wants to predict more than one disease, he/she has to go through different sites. There is no common system where one analysis can perform more than one disease prediction. Some of the models have lower accuracy which can seriously affect patients' health. When an organization wants to analyse their patient's health reports, they have to deploy many models which in turn increases the cost as well as time Some of the existing systems consider very few parameters which can yield false results.

1.3Proposed system

In multiple disease prediction, it is possible to predict more than one disease at a time. So the user doesn't need to traverse different sites in order to predict the diseases. We are taking three diseases that are Liver, Diabetes, and Heart. . As all the three diseases are correlated to each other. To implement multiple disease analyses we are going to use machine learning algorithms and Django. When the user is accessing this API, the user has to send the parameters of the disease along with the disease name. Django will invoke the corresponding model and returns the status of the patient.

2. LITERATURE REVIEW

1. According to the paper focuses about as diabetes is one of the dangerous diseases in the world , it can cause many varieties of disorders which includes blindness etc .In this paper they have used machine learning techniques to find out diabetes disease as it is easy and flexible to forecast whether the patient has illness or not . Their aim of this analysis was to invent a system that can help the patient to detect the diabetes disease of the patient with accurate results. Here they used mainly 4 main algorithms Decision Tree , Naïve Bayes , and SVM algorithms and compared their accuracy which is 85%,77%, 77.3% respectively . They also used ANN algorithm after the training process to see the reactions of the network which states whether the disease is classified properly or not . Here they compared the precision recall and F1 score support ad accuracy of all the models[1] .

2. The main aim of the paper is ,as heart plays an important role in living organisms. So the diagnosis and prediction of heart related disease should be perfect and correct because it is very crucial which can cause death cases related to heart .So Machine learning and Artificial Intelligence supports in predicting any kind of natural events .So in this paper they calculate accuracy of machine learning for predicting heart disease using k-nearest neighbor ,decision tree, linear regression and SVM by using UCI repositor dataset for training and testing . They also compared the algorithm and their accuracy SVM 83 %,Decision tree 79%,Linear regression 78%,k-nearest neighbour 87%[2].

3. The system defines that liver diseases is causing high number of deaths in India and is also considered as a life threating disease in the world. As it is difficult to detect the liver disease at early stage .So using automated program using machine learning algorithms we can detect the liver disease accurately .They used and compared SVM ,Decision Tree and Random forest algorithm and measures precision, accuracy and recall metrics for quantitative measurement. The accuracy are 95%,87%,92% respectively[3].

3. SYSTEM ANALYSIS

3.1Functional Requirement

- The system allows the patient to predict the disease
- The user adds the input for the particular disease and based on the trained model of the user input the output will be displayed .

3.2 Non Functional Requirement

- The website will provide range of the values during the prediction of the disease.
- The website should be reliable and consistent.

4. DESIGN

4.1Architecture Design

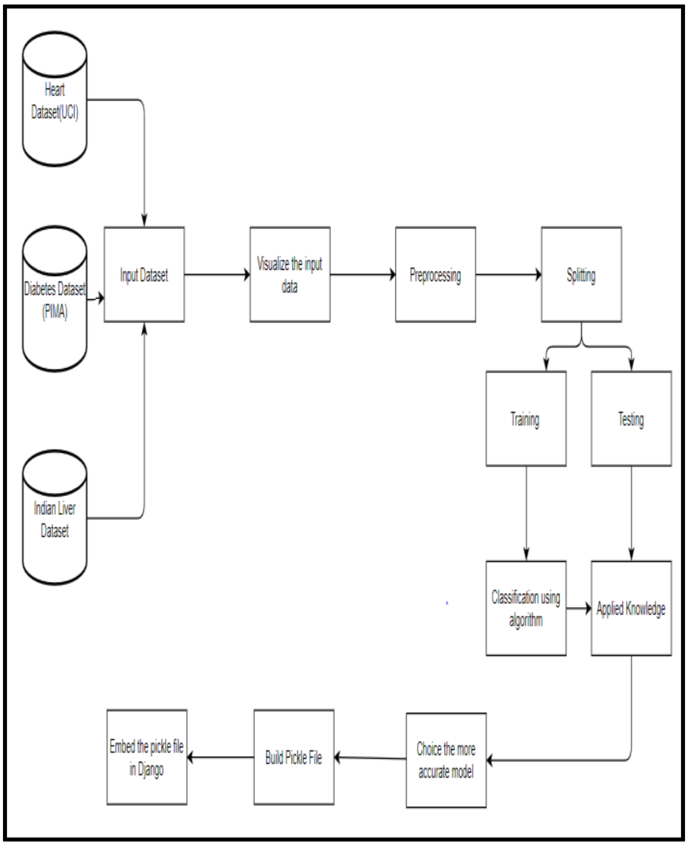
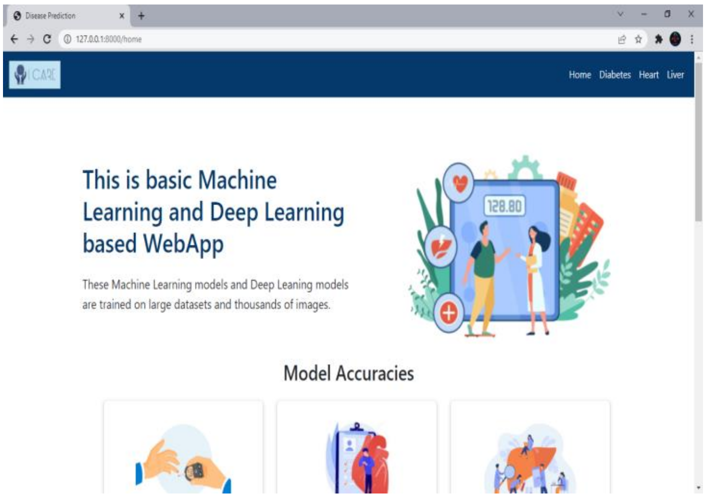


Figure No4.1: Block Diagram

In the figure no 4.1 we have experimented on three diseases that is heart,daibetes and liver as these are correlated to each other.The first step is to the dataset for heart disease,daibetes disease and liver disease we have imported the UCI dataset,PIMA dataset and Indian liver dataset respectively.Once we have imported the dataset then visualization of each inputed data takes place.After visualization pre-processing of data takes place wher we check for outliers,missing values and also scale the dataset then on the updated dataset we split the data into training and testing .Next is on the training dataset we had applied knn,xgboost and random forest algorithm and applied knowledge on the classified algorithm using testing dataset.After applying knowledge we will choose the algorithm with the best accuracy for each of the disease .Then we build a pickle file for all the disease and then integrated the pickle file with the django framework for the output of the model on the webpage.

4.2User Interface Design



This is basic Machine Learning and Deep Learning based WebApp

These Machine Learning models and Deep Leaning models are trained on large datasets and thousands of images.

Model Accuracies

Figure No4.2: Graphical User Interface

5. IMPLEMENTATION

5.1Algorithm

5.1.1. Knn Algorithm

The working of the K-NN algorithm is as followed:

- Step-1: Start to select the K value for example k=5
- Step-2: Then we will find the Euclidean distance between the points. It is calculated by the as:

$$Euclidean\ Distance = \sqrt{(X2 - X1)^2 + (Y2 - Y1)^2}$$

- Step-3: Then we will calculate the Euclidean distance of the nearest neighbour.
- Step-4: Then count the number of the data points in each category .For example found three values for Category A and two values for category B.
- Step-5: Then assign the new point to the category having maximum number of neighbours. For example Category A has highest number of neighbour so we will assign the new data point to category A.
- Step-6: So finally our Knn model is ready.

5.1.2. Random Forest Algorithm

Random Forest working is possible in two phases ,first is to create the random forest by merging N decision tree, and second is making prediction for each tree created in the first phase.

The working of the random forest is as follows:

Step-1: Firstly it will select random K data points from the training set.

Step-2: After selecting k data points then building the decision trees associated with the selected data points (Subsets).

Step-3: Then choosing the number N for decision trees that you want to build.

Step-4: Repeating step 1 and 2 .

Step-5: Finding the predictions of each decision tree, and assigning the new data points to the category that wins the majority votes.

5.1.3. XGBoost Algorithm

The working of XGBoost algorithm are as follows:

Step 1: Firstly creating a single leaf tree.

Step 2: Then for the first tree, we have to compute the average of target variable as prediction and then calculating the residuals using the desired loss function and then for subsequent trees the residuals come from prediction that was there in previous tree.

Step 3: Calculating the similarity score using formula:

$$Similarity\ Score = Gradient\frac{Gradient^2}{Hessian + \lambda}$$

where, Hessian is equal to number of residuals; Gradient2 = squared sum of residuals; λ is a regularization hyperparameter.

Step 4: Applying similarity score we select the appropriate node. The higher the similarity score more the homogeneity.

Step 5: Applying similarity score we calculate Information gain. Information gain help to find the difference between old similarity and new similarity and tells how much homogeneity is achieved by splitting the node at a given point. It is calculated by the formula:

$$Information\ Gain = Left\ Similarity + Right\ Similarity - Similarity\ for\ Roots$$

Step 6:Creating the tree of desired length using the above method pruning and regularization can be done by playing with the regularization hyperparameter.

Step 7: Then we can predict the residual values using the Decision Tree you constructed.

Step 8: The new set of residuals is calculated as:

$$New\ Residuals = Old\ Residulas + \rho\sum Predicted\ Residuals$$

where ρ is the learning rate.

Step 9:Then go back to step 1 and repeat the process for all the trees.

6. RESULT

In the system diabetes disease prediction model used knn algorithm, heart disease uses the xgboost algorithm and liver uses the random forest algorithm as these gave the best accuracy accordingly. There when the patient adds the parameter according to the disease it will show whether the patient has a disease or not according to the disease selected. The parameters will show the range of the values needed and if the value is not between the range or is not valid or is empty it will show the warning sign that add a correct value.

ACCURACY FOR EACH DISEASE:

Table No 6.1:Diabetes Disease

ALGORITHM	Diabetes
Random Forest	88%
XGBoost	89%

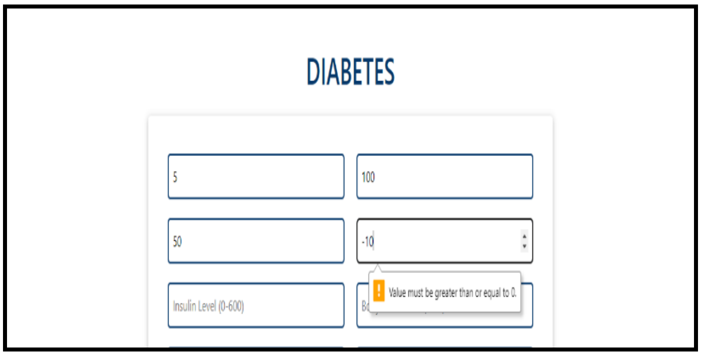
Table No 6.2:Hear Disease

ALGORITHM	Heart
KNN	85%
Random Forest	77%

Table No 6.3:Liver Disease

ALGORITHM	Liver
Random Forest	73%
XGBoost	68%

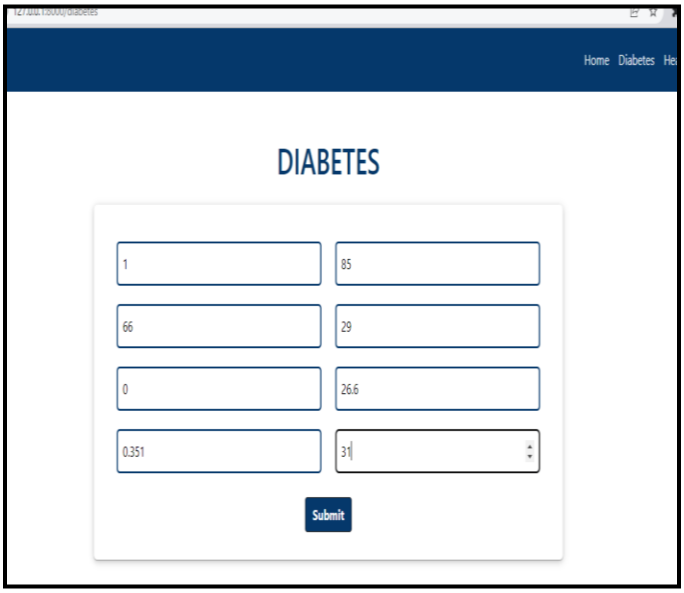
1. Error Message on inputting the incorrect value:


DIABETES

5
100
50
-10
Insulin Level (0-600)
Value must be greater than or equal to 0.

Figure No:6.1:Error Message

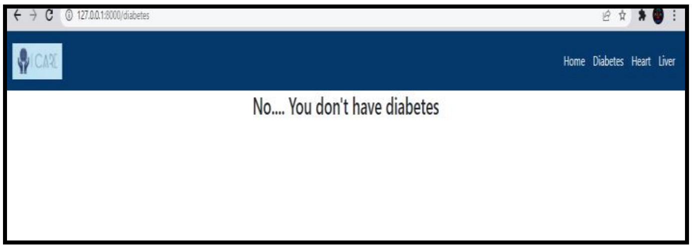
2.Diabetes Disease :



DIABETES

1
85
66
29
0
26.6
0.351
31

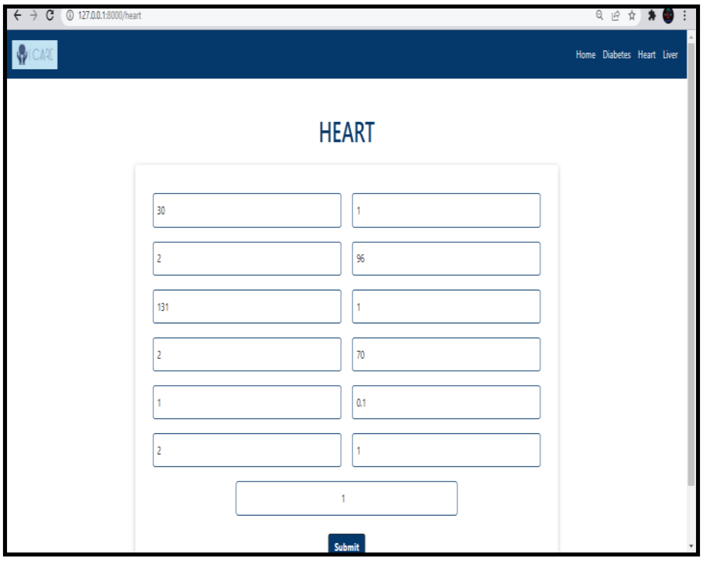
Figure No 6.2:Diabetes Disease Input Data



No.... You don't have diabetes

Figure No 6.3:Diabetes Disease Output Result

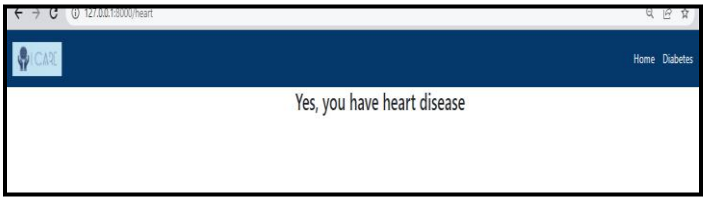
3.Heart Disease:



HEART

30
1
2
96
131
1
2
70
1
0.1
2
1
1

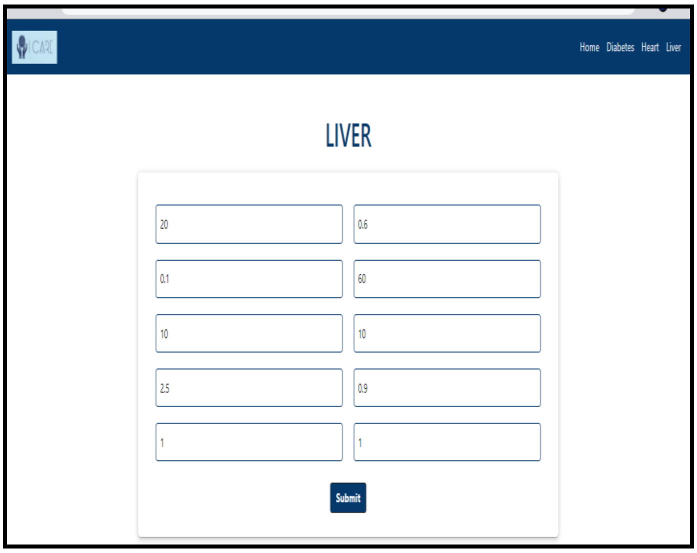
Figure No 6.4:Heart Disease Input Data



Yes, you have heart disease

Figure No 6.5:Heart Disease Output Result

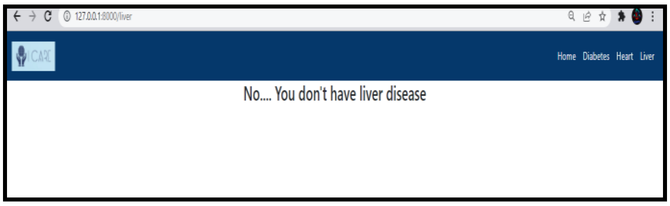
4.Liver Disease:



LIVER

20	0.6
0.1	60
10	10
2.5	0.9
1	1

Figure No 6.6:Liver Disease Input Data



No.... You don't have liver disease

Figure No 6.7:Liver Disease Output Result

7. CONCLUSION

The main objective of this project was to create a system that would predict more than one disease and do so with high accuracy. Because of this project the user doesn't need to traverse different websites which saves time as well. Diseases if predicted early can increase your life expectancy as well as save you from financial troubles. For this purpose, we have used various machine learning algorithms like Random Forest, XGBoost, and K nearest neighbor (KNN) to achieve maximum accuracy.

8. FUTURE SCOPE

- In the future we can add more diseases in the existing API.
- We can try to improve the accuracy of prediction in order to decrease the mortality rate.
- Try to make the system user-friendly and provide a chatbot for normal queries.

ACKNOWLEDGEMENT

We sincere thank to our college **Atharva College of Engineering** for giving us a platform to prepare a project on the topic "Multiple Disease Prediction System" and would like to thank our principal **Dr. Shrikant Kallurkar** for giving us the opportunities and time to conduct and research on the subject. We are sincerely grateful for **Prof. Renuka Nagpure** as our guide and **Prof. Deepali Maste,** Head of Information Technology Department, for providing help during our research, which would have seemed difficult without their motivation, constant support, and valuable suggestion. Moreover, the complication of this research paper would have been impossible without the co-operation, suggestion, and help of our friends and family.

REFERENCES

[1] Priyanka Sonar, Prof. K. JayaMalini," DIABETES PREDICTION USING DIFFERENT MACHINE LEARNING APPROACHES", 2019 IEEE ,3rd International Conference on Computing Methodologies and Communication (ICCMC)

[2] Archana Singh ,Rakesh Kumar, "Heart Disease Prediction Using Machine Learning Algorithms", 2020 IEEE, International Conference on Electrical and Electronics Engineering (ICE3)

[3] A.Sivasangari, Baddigam Jaya Krishna Reddy,Annamareddy Kiran, P.Ajitha," Diagnosis of Liver Disease using Machine Learning Models" 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)