Multiple Disease Prediction System

Aman singh
Department of Computer Science and Engineering
Chandigarh University
Gharuan, Punjab, India
amanchandansingh@gmail.com

Shobhit Srivastava
Department of Computer Science and Engineering
Chandigarh University
Gharuan, Punjab, India
sshobhit657@gmail.com

Akansha Agrawal
Department of Computer Science and Engineering
Chandigarh University
Gharuan, Punjab, India
akanshaagrawal@gmail.com

Aditya Sharma
Department of Computer Science and Engineering
Chandigarh University
Gharuan, Punjab, India
Adityasharmagreat4@gmail.com

Kirat Kaur
Department of Computer Science and Engineering
Chandigarh University
Gharuan, Punjab, India
kirat.e12999@cumail.in

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 huge data has been generated in all fields. Data in healthcare contains all information related to patients. Here, a general architecture for disease prediction in the healthcare industry has been proposed. Many existing models focus on a single disease for analysis. Like one analysis for diabetes analysis, one for cancer analysis, one for skin diseases like this. There is no common system that can analyze more than one disease at a time. Therefore, we focus on providing instant and accurate disease predictions to users about the symptoms they enter along with the predicted disease. So we design a system that is used to predict many diseases using Django. In this system we will analyze the analysis of diabetes, heart and malaria. Many other diseases may be included later. We will use machine learning algorithms and Django to implement multiple disease prediction systems. Python pickling is used to store the behavior of the model. The importance of this system analysis is that in the analysis of diseases, all the parameters that cause the disease are included, so that it is possible to detect the disease more efficiently and accurately. The behaviour of the final model will be saved as a python pickle file.

1.1 Description

Many analyze of existing systems in the healthcare industry have considered only one disease at a time. For example, one system is used to analyse diabetes, another is used to analyze diabetic retinopathy, and yet another system is used to predict heart disease. Maximum systems target a specific disease. When an organization wants to analyze the health reports of its patients, it needs to deploy many models. The approach in the existing system is useful for analyzing only specific diseases. In a multi-disease prediction system, a user can analyze more than one disease on a single web page. The user does not have to go through different places to predict whether he has a particular disease or not. In the multi-disease prediction system, the user must select the name of a specific disease, enter its parameters and click submit. The corresponding machine learning model will be invoked to predict the output and display it on the screen.

1.2 Problem system

Many existing machine learning models for healthcare analytics focus on a single disease for analysis. For example, the first one is for liver analysis, one for cancer analysis, and one for lung disease like this. If the user wants to predict more than one disease, he has to go through different pages. There is no common system where one analysis can perform more than one disease prediction. Some models have lower accuracy, which can seriously affect the health of patients. When an organization wants to analyze the health reports of its patients, it has to deploy many models, which in turn increases costs and time. Some of the existing systems take into account very few parameters, which can produce false results.

1.3 Proposed system

When predicting multiple diseases, it is possible to predict multiple diseases at once. So the user does not have to go through different places to predict the diseases. We take three diseases which are liver, diabetes and heart. Because all three diseases are related. We will use machine learning algorithms and the Streamlit web application to implement multiple disease analyses. When a user accesses this API, they must submit the disease parameters along with the disease name. Streamlit invokes the corresponding model and returns the patient state.

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 and 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 neighbour, 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 is 95%,87%,92% respectively.

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

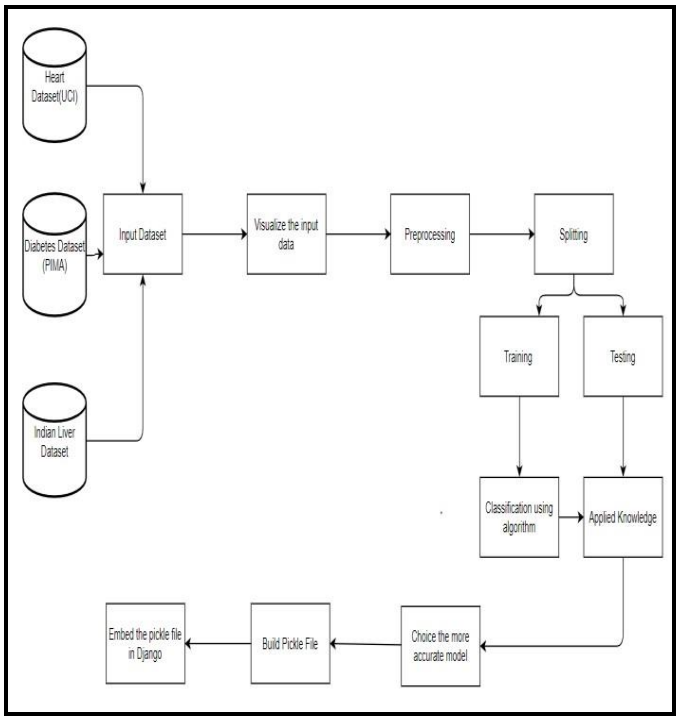


Figure No4.1: Block Diagram

In the figure no 4.1 we have experimented on three diseases that is heart, diabetes and liver as these are correlated to each other. The first step is to the dataset for heart disease, diabetes 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 inputted data takes place. After visualization pre-processing of data takes place where 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.

4.2 User Interface Design

Figure No4.2: Graphical User Interface

5. IMPLEMENTATION

5.1 Algorithm

5.1.1. Logistic Regression Algorithm

This type of statistical model (also known as *logit model*) is often used for classification and predictive analytics. Logistic regression estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables. Since the outcome is a probability, the dependent variable is bounded between 0 and 1. In logistic regression, a logit transformation is applied on the odds—that is, the probability of success divided by the probability of failure. This is also commonly known as the log odds, or the natural logarithm of odds, and this logistic function is represented by the following formulas:

Logit(pi) = 1/ (1+ exp(-pi))

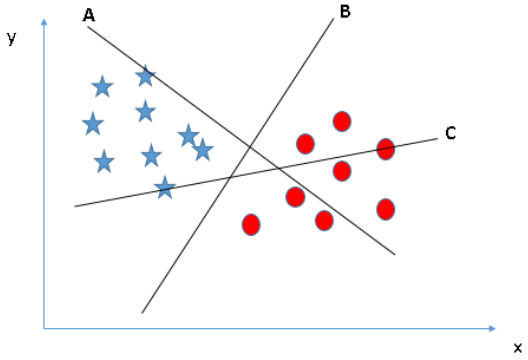
ln(pi/(1-pi)) = Beta_0 + Beta_1*X_1 + … + B_k*K_k

In this logistic regression equation, logit(pi) is the dependent or response variable and x is the independent variable. The beta parameter, or coefficient, in this model is commonly estimated via maximum likelihood estimation (MLE). This method tests different values of beta through multiple iterations to optimize for the best fit of log odds. All of these iterations produce the log likelihood function, and logistic regression seeks to maximize this function to find the best parameter estimate. Once the optimal coefficient (or coefficients if there is more than one independent variable) is found, the conditional probabilities for each observation can be calculated, logged, and summed together to yield a predicted probability. For binary classification, a probability less than .5 will predict 0 while a probability greater than 0 will predict 1. After the model has been computed, it's best practice to evaluate the how well the model predicts the dependent variable, which is called goodness of fit. The Hosmer–Leeshawn test is a popular method to assess model fit.

5.1.2. Support Vector Machine Algorithm

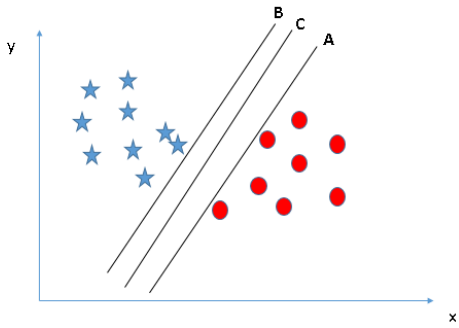
"Support Vector Machine" (SVM) is a supervised machine learning algorithm that can be used for both classification or regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is a number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well.

- **Identify the right hyper-plane (Scenario-1):** Here, we have three hyper-planes (A, B, and C). Now, identify the right hyper-plane to classify stars and circle.

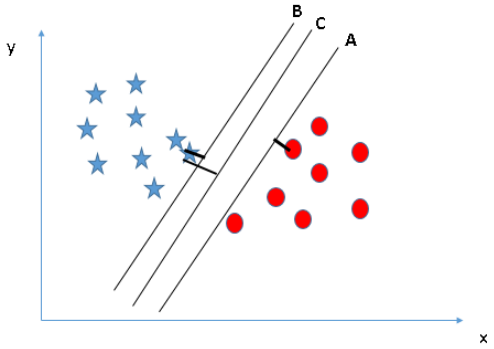


You need to remember a thumb rule to identify the right hyper-plane: "Select the hyper-plane which segregates the two classes better". In this scenario, hyper-plane "B" has excellently performed this job.

- **Identify the right hyper-plane (Scenario-2):** Here, we have three hyper-planes (A, B, and C) and all are segregating the classes well. Now, how can we identify the right hyper-plane?

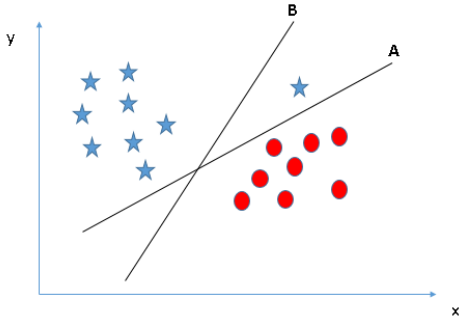


Here, maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as **Margin**. Let's look at the below snapshot:



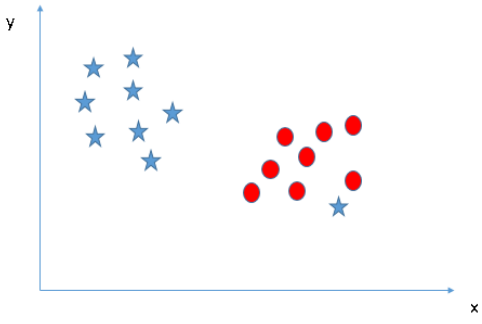
Above, you can see that the margin for hyper-plane C is high as compared to both A and B. Hence, we name the right hyper-plane as C. Another lightning reason for selecting the hyper-plane with higher margin is robustness. If we select a hyper-plane having low margin then there is high chance of miss-classification.

- **Identify the right hyper-plane (Scenario-3):** Hint: Use the rules as discussed in previous section to identify the right hyper-plane

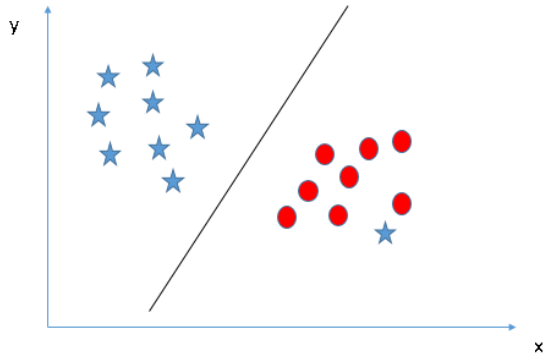


Some of you may have selected the hyper-plane **B** as it has higher margin compared to **A.** But here is the catch, SVM selects the hyper-plane which classifies the classes accurately prior to maximizing margin. Here, hyper-plane B has a classification error and A has classified all correctly. Therefore, the right hyper-plane is **A.**

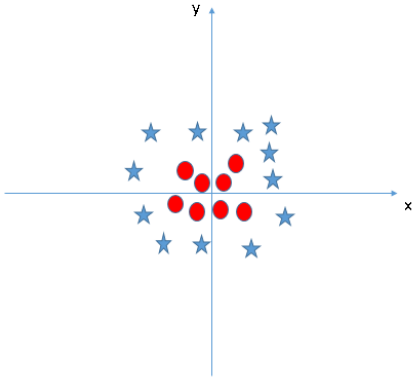
- **Can we classify two classes (Scenario-4):** Below, I am unable to segregate the two classes using a straight line, as one of the stars lies in the territory of other(circle) class as an outlier.



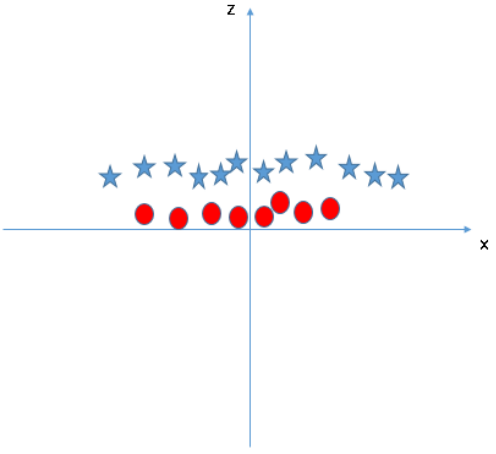
As I have already mentioned, one star at other end is like an outlier for star class. The SVM algorithm has a feature to ignore outliers and find the hyper-plane that has the maximum margin. Hence, we can say, SVM classification is robust to outliers.



- **Find the hyper-plane to segregate to classes (Scenario-5):** In the scenario below, we can't have linear hyper-plane between the two classes, so how does SVM classify these two classes? Till now, we have only looked at the linear hyper-plane.



SVM can solve this problem. Easily! It solves this problem by introducing additional feature. Here, we will add a new feature z=x^2+y^2. Now, let's plot the data points on axis x and z:



In above plot, points to consider are:

- All values for z would be positive always because z is the squared sum of both x and y
- In the original plot, red circles appear close to the origin of x and y axes, leading to lower value of z and star relatively away from the origin result to higher value of z.

In the SVM classifier, it is easy to have a linear hyper-plane between these two classes. But another burning question which arises is, should we need to add this feature manually to have a hyper-plane. No, the SVM algorithm has a technique called the **kernel trick**. The SVM kernel is a function that takes low dimensional input space and transforms it to a higher dimensional space i.e., it converts not separable problem to separable problem. It is mostly useful in non-linear separation problem. Simply put, it does some extremely complex data transformations, then finds out the process to separate the data based on the labels or outputs you've defined.

6. RESULT

In the system diabetes disease prediction model used SVM algorithm, heart disease uses the Logistic regression and Parkinson uses the SVM 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
SVM	78% (Training data)
SVM	77% (Testing data)

Table No 6.2: heart disease

ALGORITHM	HEART
Logistic regression	85% (Training data)
Logistic regression	81% (Testing data)

Table No 6.3: Parkinson's Disease

ALGORITHM	DIABETES
SVM	87% (Training data)
SVM	87% (Testing data)

Figure No 6.1: Diabetes Disease Input Data

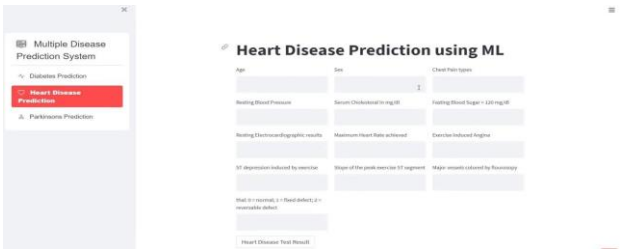


Figure No 4.2: Heart Disease Input Data

Figure No 4.3: Parkinson's Disease Input Data

7. Conclusion

The main goal of this project was to create a system that predicts multiple diseases with high accuracy. Thanks to this project, users do not need to visit various websites, which also saves them time. Early-predicted disease can extend life expectancy and save you from financial troubles. For this purpose, various machine learning algorithms such as Random Forest, XGBoost, and K next Neighbor (KNN) are used. was used to achieve maximum accuracy.

This paper offers analysis of multiple researches exhausted this field. Our planned System aims at bridging gap between Doctors and Patients which can facilitate each categories of users in achieving their goals. this technique provides support for multiple sickness prediction mistreatment completely different Machine Learning algorithms. this approach of the many systems focuses solely on automating this method that lacks in building the user's trust within the system.

By providing Doctor's recommendation in our system, we have a tendency to guarantee user's trust aspect by aspect making certain that the Doctor's won't feel that their Business is obtaining affected because of this technique.

The User interface can be improved further and a app can be created also for the same. We can also host this project

so that users can directly access the website and check their disease related queries.

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
- The User interface can be improved further and a app can be created also for the same.
- We can also host this project so that users can directly access the website and check their disease related queries.
- We can use database or real time data of users to create more efficient system.

REFERENCES

[1] Trends in coronary Heart Disease Epidemiology

[2] Center for Disease Control and Prevention (Heart Disease Facts).

[3] Asian Pacific Journal of Global Trend of Cancer Mortality rate: A 25-year study.

[4] Times Of India: Cancer cases upswing 10% in 4 years to 13.9 lakh.

[5] International Diabetes Federation: Expenditure and deaths related to diabetes.

[6] Epidemiology of Diabetes :A report of Indian Heart Association.

[7] Naveen Kishore G,V .Rajesh ,A.Vamsi Akki Reddy, K.Sumedh,T.rajesh Sai Reddy, "Prediction Of Diabetes Using Machine Learning Classification Algorithms".

[8] Gavin Pinto, Sunil Jangid, Radhika Desai, "Understanding the Lifestyle of people to identify the reasons of Diabetes using data mining".

[9] M.Marimuthu ,S.Deivarani ,R.Gayatri, "Analysis of Heart Disease Prediction using Machine Learning Techniques".

[10] Purushottam, Richa Sharma ,Dr. Kanak Saxena, "Efficient Heart Disease Prediction System".

[11] Adil Hussain She, Dr. Pawan Kumar Chaurasia," A Review on Heart Disease Prediction using Machine Learning Techniques".

[12] M. Chinna Rao ,K. Ramesh, G. Subbalakshmi,"Decision Support in Heart Disease Prediction System using Naïve Bayes".

[13] Amandeep Kaur , Jyothi Arora," Heart Disease Prediction using data mining Techniques :A survey".

[14] Noreen Fatima , Li Liu , Sha Hong, Haroon Ahmed ,"Prediction of Breast Cancer, Comparitive Review Of Machine Learning Algorithms and their analysis".

[15] Ch .Shravya ,K.Pravallika , Shaik Subhani, "Prediction of Cancer using supervised machine learning Algorithms".

[16] Nikita Rane, Jean Sunny, Rucha Kanade, Sulochana Devi," Breast Cancer classification and prediction using machine learning ".

[17] Deepti Sisodia, Dilip Singh Sisodia," Prediction of Diabetes using classification Techniques".

[18] Dr.B.Santhosh Kumar, T.Daniya, Dr. J.Ajayan," Breast Cancer Prediction using Machine Learning Algorithms".

[19] Mümine KAYA KELEŞ ,"Cancer Prediction using and Detection using Machine Learning Algorithms : A Comparitive Study".

[20] Heart Disease Dataset" by UCI.

[21] Pima Indians Diabetes Dataset" by Kaggle.