IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Health Disease Prediction using Machine Learning

Dr.N.Sreenivas **Dr.V.V.S.S.S. Balaram**

J Prasanna Krishna Reddy **T Hruthik Reddy** **M Shiva Krishna**

Abstract

Health condition prediction is a machine learning project that is used to detect different diseases that might occur in the future based on the present health condition. Most people are not aware of the symptoms of different diseases. Timely attention and proper diagnosis of different diseases will reduce the mortality rate. Different diseases might occur to an individual (like heart disease, kidney disease, lung disease, etc.). So, in this model, we use different attributes to find whether a person is prone to a particular disease or not based on different algorithms. Supervised algorithms are used for the early prediction of different diseases. Medical data sets contain a large number of features. The performance of the classifier depends on the amount of noisy data in the dataset. We select relevant features to solve this problem. We use KNN, SVM, and logistic regression classifiers for prediction. By comparing the accuracies we select the optimal algorithm to obtain results for our inputs through the user interface.

Keywords — KNN, SVM, Logistic Regression.

I. INTRODUCTION

Prediction of diseases with the help of patient treatment history and history of health data of patients and applying data mining and machine learning techniques and tools is a struggle going on for the past decades. Works have used data mining and machine learning techniques to medical data or medical profiles for prediction of specific type of diseases. These techniques tried to predict the reoccurrence of diseases. Some approaches try to do prediction on control and progression of specific diseases. The successful introduction of deep learning in dissimilar areas of machine learning has driven a change towards the use of machine learning models which can learn rich, hierarchical representations of raw data with minimum pre-processing required and produce accurate results. The progress of big data technology had created more attention to disease prediction from the perspective view of big data analysis. Various researches and experiments have been conducted by selecting the characteristics automatically from a large number of data to improve the accuracy of classification rather than the previously selected characteristics of data. According to the World Health Organization, every year 60 million deaths are occurring worldwide due to various types of health diseases. Our project aims to predict future Health Diseases by analysing data of patient's records which predicts whether they have any health disease or not using machine-learning algorithms and tools. Most researches have been conducted to describe the most influential and important factors of heath diseases as well as accurately predict the overall risk involved. Our project

aims to predict future Health Diseases by analysing data of the patients which predicts whether they have any health disease or not using machine-learning algorithms and the tools.

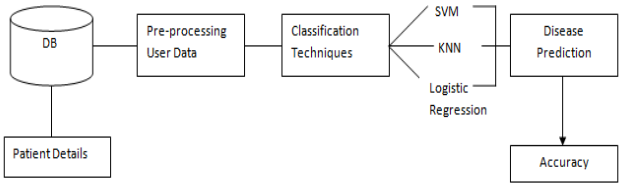
The most important part is the detection of the health diseases in a person. There are instruments and tools available where we can predict the occurrence of diseases but either there are very expensive or not efficient in calculating the chances of occurrence of diseases in humans. Fast detection of health diseases can decrease the death rate and overall heath complications of a person. But, it might not be possible to check the patients data every day in all the situations accurately and observation of a patient for 24 hours by a doctor is also not possible since it needs more necessity of doctors, time and expertise. However, we have a good amount of data in today's world which can used to analyse the data for hidden patterns by using different machine learning algorithms and tools. All the hidden patterns can be used for health diagnosis of a patient in medical field.

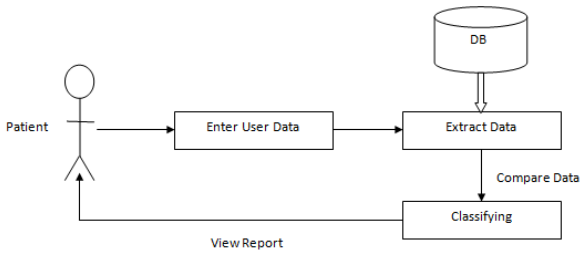
The main objective of developing this project is to use machine learning models and algorithms to predict future possibility of occurrence of health diseases by using patients datasets. To determine significant risk factors involved based on medical datasets used which may lead to prediction of health related diseases in future. To analyse and process feature and attribute selection methods and understand working principles of these methods or techniques.

II. LITERATURE SURVEY

Literature review mainly consist of three individual.sections, Firstly, finding the patient data from previous medical history and evaluating the attributes that are required for analysis. Second is about the data mining process of patients data to generate patterns that are used for understanding the reasons for having different diseases based on different health conditions. Lastly, based on these machine learning models that are used for classification of

patientss into two classes who are going to have a particular disease and the person not having that particular kind of disease.





First, we collect the existing patient data from different sources and we integrate the data into one dataset for one disease. It will be stored in a database. Pre-processing of data will be done to remove outlier's and noise from the dataset. After dataset is free from noise and outlier's classifier techniques will be used to develop the models. Here, we used three different classifier algorithms i.e., KNN, SVM and Logistic Regression. A model is developed for each classifier which is ready for prediction. We then calculate accuracy for each model and using comparative approach we compare accuracies of each model with other. Now prediction will begin from here. The patient has to enter different attributes as input and has to select through which classifier the prediction has to be done. Then after predicting the output will be displayed that whether the user is having a disease or not.

III. PROPOSED SYSTEM

Our project aims on predicting the occurrence of health diseases with maximum accuracy using some of the machine learning algorithms and classifiers. We are using K-nearest neighbours, Support vector machines and logistic regression algorithms in this project work. We enter the data through a web based interface by entering the values for the attributes given for a particular type of prediction of disease. This data input is used as test data and along with the training data we try to predict the accuracy and performance of the algorithms used. We use the classifiers which give good accuracy and are used to classify the user data. The training data to the models will be a heart disease dataset or kidney disease dataset or lung disease dataset which is already preprocessed using data analytics. We apply the KNN, SVM and logistic regression on these features and we measure the performance and accuracy of these models. The model with good accuracy classifies patients with heart Disease or without heart disease, with kidney disease or without kidney disease and finally with lung disease or without lung disease. The datasets are gathered from different resources which are already pre-processed. The user interface provides an easy-to-use visual environment and building the predictive analytics. ML process starts from a pre-processing data stage followed by feature selection or pattern selection based on data cleaning, data classification and finally performance evaluation. KNN, SVM, Logistic Regression are used to predict the health conditions of a person. Through the outcome of this project, a person can take necessary precautions like consulting a doctor or taking medicines or maintaining a proper diet to keep his health good

DATASETS

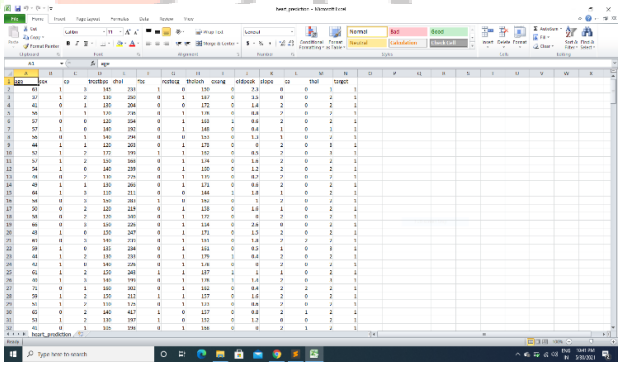
The source of Dataset is Kaggle

There are three Datasets used in predicting three different diseases. They are:

- Heart Disease Dataset
- Kidney Disease Dataset
- Lung Disease Dataset

Attributes of Heart dataset:

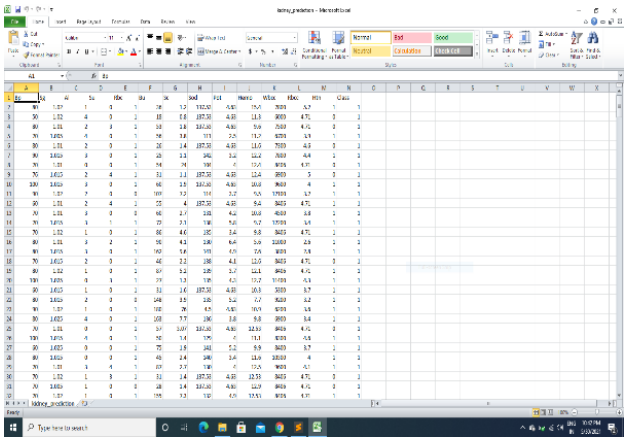
age - age in years
gender
cp - chest pain type
trestbps - resting blood pressure
chol - serum cholesterol in mg/dl
fbs - fasting blood sugar
restecg - resting electrocardiographic results
thalach - maximum heart rate achieved
exang - exercise induced angina
oldpeak - ST depression induced by exercise relative to rest looks at stress of heart during exercise. unhealthy heart will stress more
slope - the slope of the peak exercise ST segment
ca - number of major vessels (0-3) coloured by fluoroscopy
thal - thallium stress result
target - have disease or not



Attributes of Kidney dataset :

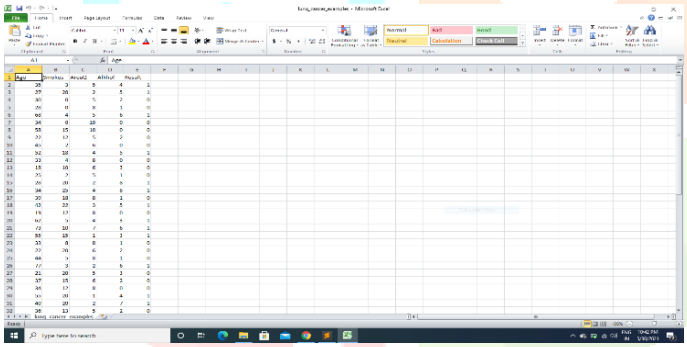
Bp - Blood Pressure
Sg - Specific Gravity
Al - Amyloidosis
Su - Sugar Levels
Rbc- Red Blood Cells
Bu - Blood level
Sc - Sickle Cell
Sod - superoxide dismutase
Pot - Potassium Nitrate
Hemo - Haemoglobin level
Wbcc- White Blood cells count

Rbcc- Red Blood cells count
Htn- Hypertension
Class- have disease or not



Attributes of Lungs dataset:

Age - Age of the patient in years
Smokes - It describes number of cigarettes that the patients consumes in a day
AreaQ - Lungs Area
Alcohol - It is the amount of alcohol that the patient consumes in a day



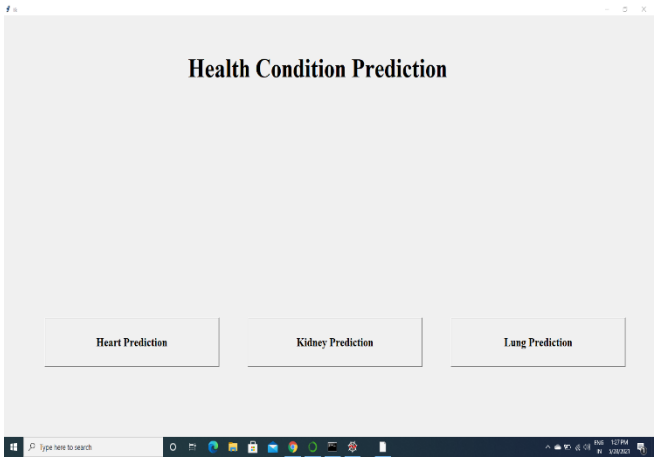
PROPOSED ALGORITHM:

a. K-Nearest Neighbours
b. Support Vector Machine
c. Logistic Regression

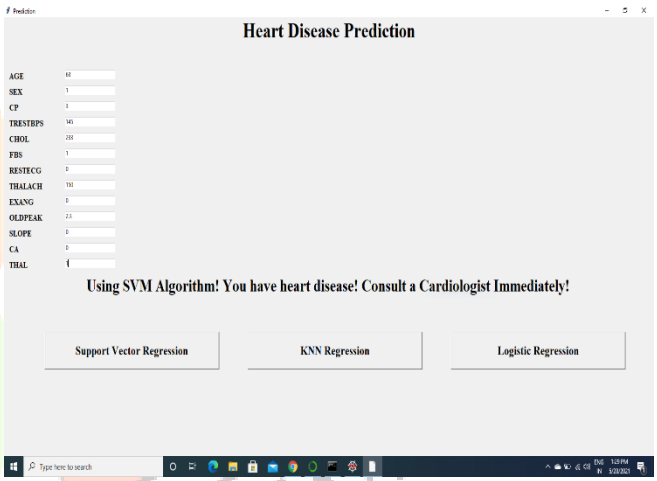
IV. RESULTS

The project aims to predict that whether a person has any likelihood of getting any kind of diseases .

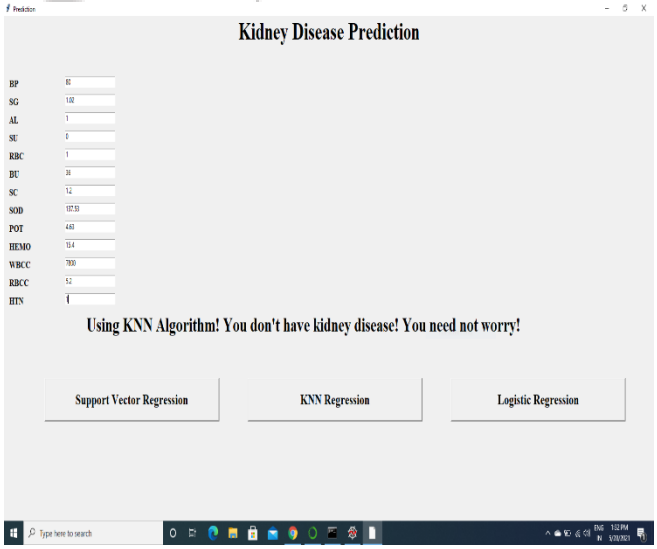
HOME PAGE



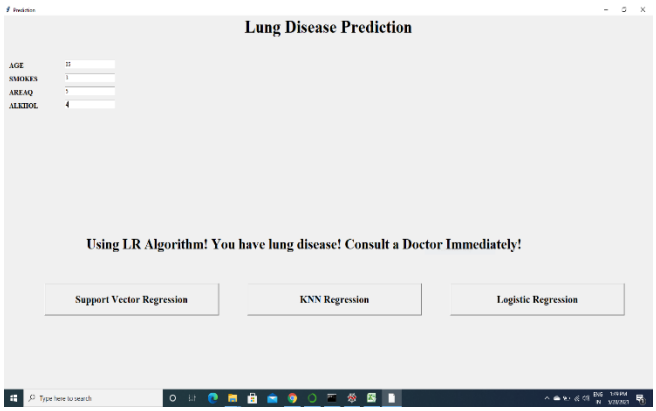
HEART DISEASE PREDICTION USING SVM



KIDNEY DISEASE PREDICTION USING KNN



LUNG DISEASE PREDICTION USING LOGISTIC REGRESSION



IV. CONCLUSION

We proposed a method for health condition disease prediction using machine learning techniques; these results showed a great accuracy standard for producing a better estimation result. By introducing new proposed KNN, SVM and logistic regression classification, we find the problem of prediction rate without equipment and propose an approach to estimate the heart rate and condition. Sample results of health values are to be taken at different stages of the same subjects; we find the information from the above input via ML Techniques. Firstly, we introduced a support vector classifier based on datasets. Same as we proposed 3 classifications for Heart, Kidney and lung diseases, these results showed a great accuracy standard for producing a better estimation result.

REFERENCES

1. Sana Bharti,2015 .Analytical study of heart disease prediction comparing with different algorithms; International conference on computing, communication and automation(ICCA2015).

2. Monika Gandhi,2015. Prediction in heart disease using techniques of data mining, International conference on futuristic trend in computational analysis and knowledge management(ABLAZE- 2015)

3. S.Ramya, Dr. N.Radha, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms," Proc. International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 1, January 2016.

4. S.Dilli Arasu and Dr. R.Thirumalaiselvi, “Review of Chronic Kidney Disease based on Data Mining Techniques”, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp. 13498-13505

5. Monsi, J., Saji, J., Vinod, K., Joy, L. and Mathew, J.J., 2019. XRAY AI: Lung Disease Prediction Using Machine Learning. International Journal of Information.

6. Ausawalaithong, W., Thirach, A., Marukatat, S. and Wilaiprasitporn, T., 2018, November. Automatic lung cancer prediction from chest X-ray images using the deep learning approach.