HEALTH RISK ESTIMATOR USING MACHINE LEARNING TECHNIQUES

[1]Erum Parkar, [2]Hamira Shaikh, [3]Taskeen Merchant, [4]Sonali Suryawanshi

[1]Student, [2]Student, [3]Student, [4]Professor,
[1]Computer Engineering Department,
[1]Rizvi College of Engineering, Mumbai, India.

Abstract: Prediction of critical diseases and Machine Learning go hand in hand as the predictive models use Machine Learning algorithms. But accurate prediction on the basis of symptoms becomes too difficult for doctors. The most challenging task is the accurate and correct prediction of the disease. There are instances where online medical help or healthcare advice is easier or faster to grasp than real world help. People often feel reluctant to go to hospital or physician for minor symptoms. However, in many cases, these minor symptoms may trigger major health hazards. This project is based on creation of a web application for detecting fatal diseases such as Heart disease, Diabetes using Machine Learning The project implements classification models such as K Nearest Neighbor (KNN) and Random Forest Classifier. The values and symptoms on various parameters entered by the user will enable the model to predict if a person is at a risk of disease. The system will provide the results based on the features extracted. It also saves a lot of time for the patients or doctors and they can further proceed for treatments or other procedures faster.

Index Terms **- Machine Learning, K Nearest Neighbor (KNN) Classifier, Random Forest Classifier.**

I. INTRODUCTION

The rates of incurable diseases are increasing day by day due to environmental conditions, stress, weak immunity and so on. From the past few years, a sudden growth in death rates due to heart disease and diabetes has been observed. It is becoming difficult for individuals to tackle these diseases at the last stage. Therefore, an early prediction of these diseases is essential and which is now considered as an important task. Medical sector focuses on large amount of data every hour. Due to large volume of data from different sources, it becomes a challenging task to find data analysis which can help to predict accurate results for patient. The growth of machine learning in healthcare and its ability to process large datasets and derive meaningful insight from it helps the doctors in providing planning, treatment and care to the patients. By taking inspiration from machine learning technology, we will propose a system which can aid early disease prediction based on data analysis.

In this project we will be using some of the powerful machine learning algorithms to predict the risk of heart disease and diabetes with the help of the data and the symptoms. We will first understand about diabetes prediction system. Diabetes is a genetic disease and factors such as excess of sugar in blood, overweight can develop diabetes in this paper, parameters used within the facts set to locate the diabetes are Glucose, Blood pressure, pores and skin thickness, Insulin, Age. Given a set of inputs are BMI (Body Mass Index), BP (Blood Pressure), Glucose Level, Insulin Level based on these features it predicts whether you have diabetes or not. Another fatal disease i.e., Heart Disease. Heart disease is one of the most common and dangerous diseases nowadays, and an early diagnosis of such a disease is a crucial task for many health care providers to prevent their patients for such a disease and to save lives. The diagnosis of heart disease is usually based on signs, symptoms and physical examination of the patient. There are several factors that increase the risk of heart disease, such as smoking habit, body cholesterol level, family history of heart disease, obesity, high blood pressure, and lack of physical exercise. For early prediction of both the disease the machine learning technique we are using is KNN Classifier and Random Forest Classifier. The model is trained using KNN Classifier and Random Forest Classifier to predict the positive or negative result. This study proposes heart disease and diabetes prediction using KNN and Random Forest Classifier with instant measurement parameters. KNN and Random Forest Classifier are one of the top data mining algorithms which frequently used in disease prediction method. The end result will be on web application which allows users that can be a patient or a doctor to access them and they can predict their risk of disease based on the symptoms entered by them. The system can answer the complex queries for diagnosing heart disease and diabetes; therefore, it can be helpful to health care practitioners to make intelligent clinical decisions.

II. LITERATURE REVIEW

In [1], the aim of this experiment is to implement machine learning algorithms to improve computer supported diagnosis reliability for diabetes. While there are many papers on this, this paper is different. Rather than implementing one machine learning algorithm or comparing different machine learning algorithms, this paper has focused on joint implementation of two machine learning algorithms i.e., Support Vector Machine (SVM) and Naïve Bayes Classifier. The paper describes a joint MATLAB implementation of the Support Vector machine and Naïve Bayes methods in a new dataset acquired from the patients examined for diabetes in Kosovo. This has been done to improve the reliability of the decision by using the power of both algorithms in minimizing their individual weakness. The results show the mean value of the SVM classifier performance - accuracy of 95, 52 % while for the Naïve Bayes classifier the classifier accuracy is 94, 52%. Both values vary in 1% of classification performance margin during various iterations. This also shows the high stability of the classifier. The conclusion of this experiment showed us that the joint implementation of the two machine learning algorithms improves the overall reliability of the system outcome significantly, which is crucial in the computer-supported diabetes diagnostic process.

In [2], the main aim of this project is to predict heart disease by providing symptoms and if the disease is detected at an early stage, then preventative measures can be taken and hence mortality rate can be controlled. The experimental workflow shows that first the data is pre-processed afterwards feature selection is done followed by classifications/predictions and the performance is evaluated to give the required result. The prediction model in this paper has different features and several classifications techniques. Various classifications techniques were used such as Decision Trees, Language Model, Support Vector Machine, Random Forest, Neural Network, K-Nearest Neighbor, Naïve Bayes out of which Hybrid Random Forest with Linear Model (HRFLM) technique

is introduced because it has best accuracy. HRFLM method uses all features without any restrictions of feature selection. Here, 70% of the data is used for training and the remaining 30% is used for classification. The dataset which is used has a record of 303 patient records, where 6 records are with some missing values. There is total 13 different attributes in this dataset out which 2 attributes sex and age are personal details while the rest 11 attributes are important. This paper gives us the brief idea of how we can figure out the algorithm which we will use in our project.

In [3], the aim of this experiment is to develop a system which might predict the diabetic risk level of a patient with a better accuracy. Various information mining algorithms present different decision support systems for assisting health specialists. The effectiveness of the decision support system is recognized by its accuracy. Therefore, the objective is to build a decision support system to predict and diagnose a certain disease with extreme precision. The authors have compared different algorithms like Decision Tree, Artificial Neural Network (ANN), Naïve Bayes and Support Vector Machine (SVM), for their accuracies to predict the disease. They have used parameters like precision value, recall value and F1-Score. T

In [4], In today's HealthCare Industry, a vast amount of data is driven every hour, and it is difficult to manage them. In such scenarios data analytics can help derive insights on systemic wastes of resources, can track individual practitioner performance, and can even track the health of populations and identify people at risk for chronic diseases. But on the other hand, due to lack of precision in data quality, the accuracy of analysis decreases. This makes the task of selecting features more difficult. The data features have a major influence on our model. Irrelevant or partially correct features can prove to be a disadvantage for the performance. The proposed method was using feature selection technique, but they faced difficulties due to improper data quality. To resolve this issue, an Improved Ant Colony Optimization based Feature Selection (IACO) algorithm is presented

III. Problem Statement and Objective

Problem Statement: By using Machine Learning, we are going to develop a system which can aid early disease prediction based on data analysis. The two major diseases - Diabetes and Heart Disease will be predicted using several machine learning algorithms.

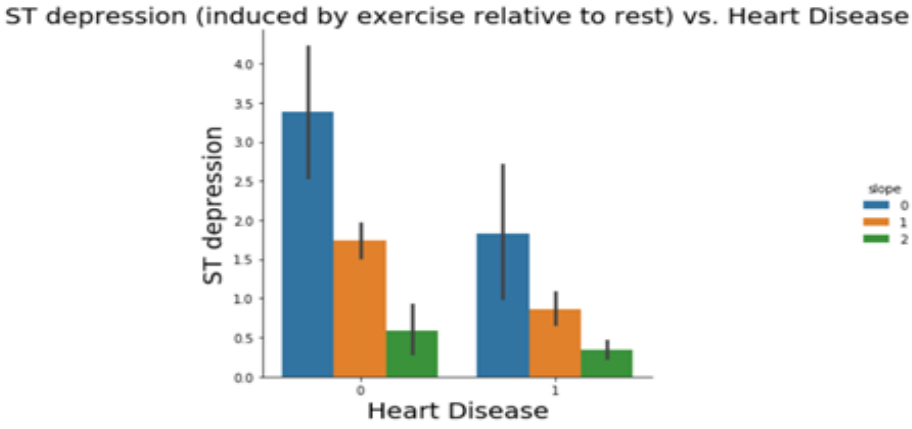
Objective: 1. Develop a Machine Learning model using KNN and Random Forest Classifier.
2. Train the model to predict the diseases correctly.
3. Deploy the model using the Flask.

IV. Proposed System

3.1 Heart Disease Prediction

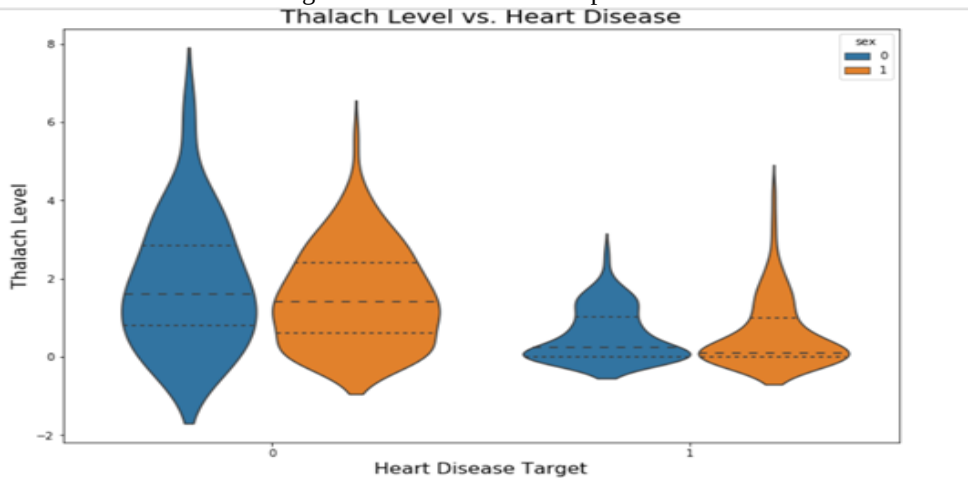
First, we import different libraries. Then we import the csv file in code and display it in which we print a total no of rows and columns. There are a total of 303 rows and 14 columns. We print unique values of the attributes. There are 41 unique values for attribute age, 2 values for sex… hence so on. We summarize the data to know the count, mean, standard deviation, minimum and maximum values for particular attributes. As per observation in code, we don't have any missing values to deal with. Target value with 1 describes a person is suffering from heart disease. Target value with 0 describes the person is not suffering from heart disease. As we count how many 1s and 0s are present in the target attribute, there are a total 165 1s and 138 0s. That means the proportion between our positive and negative results is almost equal. We can see there is a positive correlation between chest pain (cp) & target (our predictor). This makes sense since, the greater amount of chest pain results in a greater chance of having heart disease. Cp (chest pain), is an ordinal feature with 4 values: Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic. In addition, we see a negative correlation between exercise induced angina (exang) & our predictor. This makes sense because when you exercise, your heart requires more blood, but narrowed arteries slow down blood flow.

Figure 1. Exercise to Rest Comparison Graph



We have also seen data using pair plot for attributes like age, trestbps, chol, thalac, oldpeak. When the ventricle is at rest St segment depression occurs. If the St segment is below baseline there are chances of heart disease. In the violin plot the y axis is for old peak value and x describes target output.

Figure 2. Thalach Level Comparison



Different hues determine gender. Now we will try to filter the data. We will count, find mean, max, min values for people with target 1 that is those who suffer from heart disease. Same with target 0 that is those who do not suffer from heart disease. We will then split the data into training and testing sets. Now using different supervised learning models, we will train our model and determine which model has best accuracy. The best accuracy from all is the Random Forest Classifier. Now in the last line of code we provide the input values and based on the input values our model will give out 1 if heart disease is present or value 0 if heart disease is not present.

3.2 Diabetes Detection

For diabetes detection, the proposed system is to take an openly available dataset which was downloaded from Kaggle. The dataset contains 768 patients with 9 attributes. Table 1 describes the attributes that were used.

Table 1. Description of Attributes

Attribute	Description
preg	No. of pregnancies (Numeric)
plas	Plasma glucose concentration (Numeric)
pres	Diastolic blood pressure (mm Hg) (Numeric).
skin	Triceps skin fold thickness (mm) (Numeric).
insu	2-Hour serum insulin (mu U/ml) (Numeric).
mass	Body mass index (Numeric).
pedi	Diabetes pedigree function (Numeric)
age	Age (Numeric).
class	Diabetes or not diabetes (0/1).

The main objective of this section of the project is to determine the percentage of risk of an individual to get diabetes. After the dataset is chosen and understood the next step is to use data mining or a machine learning algorithm to train and test the machine so as to properly predict diabetes. Machine learning is a part of Artificial Intelligence which deals with the computer algorithms that improve automatically through experience. There are two types of machine learning algorithms: Supervised learning & Unsupervised learning. Supervised learning means that the machine is pre-trained using labeled data, whereas, Unsupervised learning means that the machine tries to find the pattern of the inputs and outputs on its own. For this project we have used a Supervised learning algorithm i.e., KNN (K-Nearest Neighbors).

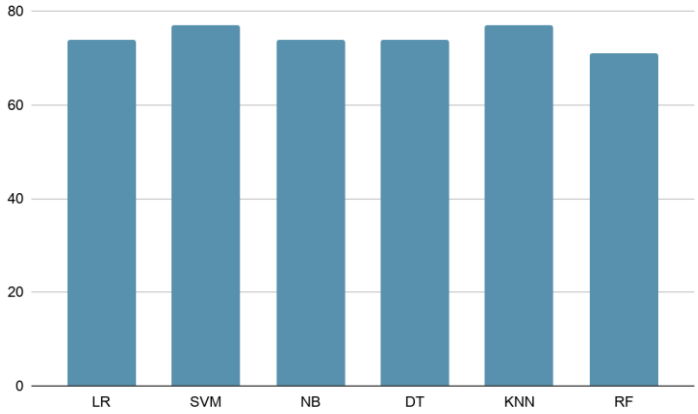
As soon as the algorithm is selected, we make a machine learning model using python and its libraries. We divide the dataset into training data and testing data, and prepare the model which will deal with the prediction. After the model is ready, we convert the python file to a PICKLE file i.e., a file with the extension (.pkl). The PICKLE is a python module which allows objects to be serialized to files on disk and deserialized back into the program at runtime. We do this to save bytes stream as well as it makes it easier to make a webapp using Flask which is a web application framework written in python.

After the PICKLE file is created, we start with the webapp. The first thing to do is to create a web template for diabetes detection using HTML, CSS and BOOTSTRAP. This web template will include the form in which the user will enter the data for prediction. As soon as the template is ready, we start creating a flask application. This application will take values from the form in the web template, unpickle the PICKLE file, and predict the results using the former two. Then it will send the result to the web template which will display the results on the screen. When we run the flask application in the IDE, we get a URL that will lead to the web template and the prediction process is started.

V. Results And Discussion

In this project, for diabetes only one algorithm was used which is KNN Classifier. This was due to intense research using various IEEE papers and reference papers. During the research it was found that out of the six algorithms chosen KNN and SVM had the highest accuracy result.

Figure 3. Accuracy of Algorithms



The following figure is the confusion matrix and the accuracy of the model which was created using KNN.

Figure 4. Model Accuracy & Confusion Matrix

```
684      136            82  ...   69       0

[576 rows x 8 columns]
The confusion matrix is
 [[130   0]
 [  1  61]]
The accuracy score is
 0.9947916666666666

Process finished with exit code 0
```

In this project for heart disease prediction, we have used 4 models for testing and training.

Table 2. Accuracy Score of algorithms

Models	Accuracy Score
K Neighbors Classifier	67.21%
Support Vector Classifier	81.97%
Decision Tree Classifier	81.97%
Random Forest Classifier	95..08%

Random Forest Classifier has the best accuracy score. Hence, we have used Random Forest Classifier for the deployment of Heart disease prediction.

REFERENCES

[1] P. Sonar and K. JayaMalini, "Diabetes Prediction Using Different Machine Learning Approaches," 2019 3rd International Conference on Computing Methodologies and
Communication (ICCMC), Erode, India, 2019, pp. 367-371, doi: 0.1109/ICCMC.2019.8819841.

[2] M. A. Sarwar, N. Kamal, W. Hamid, and M. A. Shah, "Prediction of Diabetes Using Machine Learning Algorithms in Healthcare," 2018 24th International Conference
on Automation and Computing (ICAC), Newcastle upon Tyne, United Kingdom, 2018, pp. 1-6, doi: 10.23919/IConAC.2018.8748992.

[3] Z. Tafa, N. Pervetica and B. Karahoda, "An intelligent system for diabetes prediction," 2015 4th Mediterranean Conference on Embedded Computing (MECO),
Budva, 2015, pp. 378-382, doi: 10.1109/MECO.2015.7181948.

[4] G. G. Ladha and R. Kumar Singh Pippal, "A computation analysis to predict diabetes based on data mining: A review," 2018 3rd International Conference on
Communication and Electronics Systems (ICCES), Coimbatore, India, 2018, pp. 6-10, doi: 10.1109/CESYS.2018.8724016.

[5] Designing Disease Prediction Model Using Machine Learning Approach. Published in: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC).

[6] Disease Prediction Using Machine Learning Akash C. Jamgade, Prof. S. D. Zade Student, Dept. of Computer Science and Engineering, Priyadarshini Institute of Engineering & Technology, Nagpur, Maharashtra, India .International Research Journal of Engineering and Technology (IRJET)

[7] Study of machine learning algorithms for special disease prediction using principal of component analysis,December 2016
Conference: 2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)