An Approach to detect multiple diseases using machine learning algorithm

An Approach to detect multiple diseases using machine learning algorithm

Indukuri Mohit[1], K.Santhosh Kumar[2], Avula Uday Kumar Reddy[3], Badhagouni Suresh Kumar[4]
[1]Vardhaman College of Engineering, Hyderabad, India

E-mail: mohitindukuri100@gmail.com

Abstract. There are multiple techniques in machine learning that can in a variety of industries, do predictive analytics on large amounts of data. Predictive analytics in healthcare is a difficult endeavour, but it can eventually assist practitioners in making timely decisions regarding patients' health and treatment based on massive data. Diseases like Breast cancer, diabetes, and heart-related diseases are causing many deaths globally but most of these deaths are due to the lack of timely check-ups of the diseases. The above problem occurs due to a lack of medical infrastructure and a low ratio of doctors to the population. The statistics clearly show the same, WHO recommended, the ratio of doctors to patients is 1:1000 whereas India's doctor-to-population ratio is 1:1456, this indicates the shortage of doctors. The diseases related to heart, cancer, and diabetes can cause a potential threat to mankind, if not found early. Therefore, early recognition and diagnosis of these diseases can save a lot of lives. This work is all about predicting diseases that are harmful using machine learning classification algorithms. In this work, breast cancer, heart, and diabetes are included. To make this work seamless and usable by the mass public, our team made a medical test web application that makes predictions about various diseases using the concept of machine learning. In this work, our aim to develop a disease-predicting web app that uses the concept of machine learning-based predictions about various diseases like Breast cancer, Diabetes, and Heart diseases.

Keywords: Logistic Regression, Support Vector Machine (SVM), K-Nearest -Neighbors (KNN), Diabetes, Breast cancer, Heart diseases.

1.Introduction

Breast cancer, diabetes, and heart disease are for the most part driving reasons for death in the present society. Heart disease is a term that refers to a group of illnesses that affect your heart. Arrhythmias (issues with heart rhythm), coronary artery disease, and congenital heart defects are all diseases that fall under the category of heart illness (the defects of the heart you are born with).The term "heart disease" is often used instead of the term "cardiovascular disease."The cardiovascular disease normally indicates heart attack, angina (heart pain), or stroke, also conditions that affect your rhythm valves or muscles of your heart also referred to as heart diseases.[1]An article published in "Science Direct" states that in India the annual deaths are nearly 11 million, of which 28 percent of deaths are due to cardiovascular disease.[2]According to the "Centers for Disease Control and Prevention," a death from cardiovascular disease happens every 36 seconds in the United States. Breast cancer is a form of cancer that develops in the breast tissue. A change in the contour of the breast is one of the signs of breast cancer and also a lump in the breast, fluid eruption from the nipple, or reddish-pink or scaly patches of skin.[3]One of the highest mortality rates recorded across the globe was due to cancer. In 2015, 87 lakh deaths are caused only by cancer. In the deaths caused, breast cancer is one of the highest mortality rates recorded right behind lung cancer.[4]The article published by "Times of India" on 19th August 2020 states that around

2.4 lakh breast cancer cases are to be expected to be the most common site in India by 2025. Diabetes is a type of disease that occurs when your glucose level in the blood, also called blood sugar is very high. Blood sugar is your major energy source that comes from the food you consume. Insulin is a hormone released by the pancreas that extracts glucose from food, which helps to run metabolic activities.[5] According to the "International Diabetes Federation" across the globe, there are 42 lakhs of deaths caused by diabetes, and around 760 billion dollars USD are spent on diabetes (as a part of health expenditure).[6]In India, over 10 lakh people die annually due to diabetes (Epidemiology of Diabetes), and according to the "Indian heart Association" nearly 11 crore individuals will end up suffering from diabetes by 2035.

The past existing AI models for medical care examination are centered around one sickness for each investigation. Like one examination if for diabetes investigation, one for malignancy examination, one for skin infections like that. There is no regular framework where one investigation can perform more than one infection expectation. In our proposed system, we unify multiple diseases under a single user interface where you can perform predictions on Heart diseases, breast cancer, and diabetes. In this work, we are using the machine learning classification algorithms like LogisticRegression, Support Vector Machine (SVM), K-Nearest-Neighbors (KNN) to perform the prediction of multiple diseases.

2.Relevent Work

This section describes the study of previously proposed models for predicting the diseases which are related to our proposed work. Several studies have been made for detecting various diseases. They have applied various data mining techniques for efficiently predicting a variety of diseases.[7]G Naveen Kishore and few other authors proposed the work named Prediction Of Diabetes Using Machine Learning Classification Techniques proposed. In this work, various classification algorithms like SVM, Logistic Regression, Decision Tree, KNN, Random Forest are utilized on the 769 instances of the Pima dataset which contain features like Pregnancies, Blood pressure, body mass index, etc. They have reported the highest accuracy as 74.4 %for the classification algorithm Random Forest and the lowest accuracy in this work is attained by the KNN reported as 71.3%. [8]The work "Understanding the lifestyle of people to identify the reasons for Diabetes using data mining" proposed by Gavin Pinto, Radhika Desai, and Sunil Jangid discussed reducing the risk of diabetes disease using data mining techniques and also discussed diabetes sub-classification. The authors used Naïve Bayes and SVM classification algorithms on the dataset collected by a survey using google forms and reported the accuracy of 64.92 for SVM and 60.44 for Naïve Bayes.

[9]In the work presented by M.Marimuthu, S.DeivaRani, Gayatri. R described the cardio diseases in a detailed manner and also applied the classification algorithms like SVM, Decision Tree, Naïve Bayes, K-Nearest Neighbors on the Framingham dataset from Kaggle. The authors compared various machine learning algorithms for the forecast of the risk of heart disease. The highest reported accuracy in this work is 83.60% for the KNN classification algorithm. [10]In the work proposed by Purushottam, Richa Sharma and Dr. Kanak Saxena discuss cardiovascular sickness by using the implementation of Knowledge Extraction based on Evolutionary Learning (java programming technique for making the development model for data mining issues). The highest reported accuracy in this work is 86.7%. [11][12]M. Chinna Rao, K. Ramesh, and G. Subbalakshmi presented a decision support system for heart disease prediction utilising the Nave Bayes classification method, which discussed the extraction of hidden information heart disease dataset that can address complex queries.[13]Amandeep Kaur and Jyothi Arora presented a study that covered the examination of algorithms such as KNN, SVM, ANN, and Decision Tree on the heart disease dataset and plotted the accuracies graph.[14]Noreen Fatima proposed work on the Cancer forecast the data mining techniques and machine learning techniques that can predict cancer effectively on the large health records and described the study previous existing models.

[15]Ch. Shravya, K.Pravallika, Shaik Subhani presented the work on Breast cancer prediction using Supervised machine learning techniques on the dataset and also analyzed the results with (PCA)principal component analysis and also used the dimensionality reduction and explained in a well-mannered way.[16]Nikitha Rane, Jean Sunny presented work on the classification of Cancer using machine learning concepts and their major discussion point is detecting cancer in very early stages so that a lot of lives can be saved.[17]Dilip Singh Sisodia ,Deepti Sisodia predicted diabetes using classification techniques and reported an accuracy of around 76% on the Pima dataset.[18] Dr.

J.Ajayan, Dr.B.Santhosh Kumar ,T.Daniya have predicted the occurrence of Breast cancer using KNN algorithm with accuracy value of 83.33%.[19]Mümine KAYA KELEŞ predicted cancer usingRandom Forest algorithms and reported an accuracy value of 92.20%.

3. Methodologies

In this section consists of the methodology adopted by our proposed work. As stated earlier our work aims to develop a web application to detect diseases like breast cancer, diabetes, and heart diseases using machine learning models. The machine learning methodologies used in our proposed work are as follows:

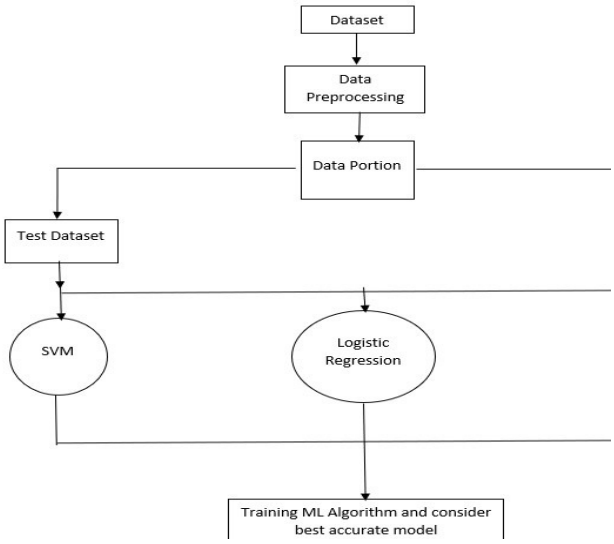

Figure 1: Process Work Flow

3 .1 Logistic Regression

Logistic regression algorithm uses the logistic function, so this algorithm is named Logistic Regression. The logistic function is an "S" shaped curve developed for statistical functionalities and the curve is plotted between 0 and 1. For the representation purpose logistic regression uses equations like linear regression.

Logistic regression Equation

Y=1/(1+EXPO(-value)) –(1)

Input values (generally termed as x) and co-efficients(Beta) are linearly combined to predict the value of output(termed as y).

Logistic regreesion Equation

y=EXPO(u0+u1*x)/(1+EXPO(u0+u1*x)). –(2)

y is predicted outcome, a0 is intercept or bias, and a1 is single input coefficient value.

The models of logistic regression predict the probability of first-class (or can be termed as default class). For example, if we are building a model for predicting the gender of a person using the height and the default class might be male which can be formally written as

P(gender=male|height). –(3)

For prediction probabilities must turn into binary values, either 0 or 1.Probabilities are turned into predictions by using the logistic function. The model can be composed as

y=EXPO(u0+u1*x)/(1+EXPO(u0+u1*x)). –(4)

Further solving we get the equation as:

ln(p(x)/1-p(x)) =u0+u1*x.–(5)

The left-hand size equation(ratio) is called odds of first-class or default class. The odds are calculated as the probability of an event divided by the probability of its complement event.

ln(odds)=u0+u1*x. –(6)
Predictions with logistic regression are quite easy and simple to implement. For example, let us assume finding the gender based on the height of a person, let us consider the height as 150, and assuming coefficients u0=-100 and u1=0.6.
y = EXPO(u0 + u1*X) / (1 + EXPO(u0 + u1*X)) -(7)
y = EXPO(-100 + 0.6*150) / (1 + EXPO(-100 + 0.6*150)) –(8)
y = 0.0000453 –(9)
The likelihood that got almost zero is male. Continuously practice, we utilize the specific probabilities.

3 .2 K-Nearest Neighbors (KNN):

3 .2.1 Introduction
KNN is a machine learning technique utilized for regression as well as classification. The algorithm is considered computationally expensive because it involves multiple iterations to get the best accuracy possible. This technique is a supervised machine learning technique which means that the data is labeled and the algorithm learns to predict the output from the input data. The algorithm also works fine even if the training data is large and contains noisy values.
The algorithm divides the dataset into test and training datasets. The training dataset is used for model building and training. The test data is predicted based on the model built. Presently we figure the distance between the prepared k-nearest element esteems and test point.
3 .2.2 Distance Metrics
Several distance measures are used to find the distance between test inputs and data feature value.
S represents the distance metric
1.Minkowski Distance:

$$S(i,j) = \sqrt[q]{|ui1 - uj1|^q + |ui2 - uj2|^q + \cdots + |uik - ujk|^q} \quad \text{–(10)}$$

2.Euclidean Distance:
q=2

$$S(i,j) - \sqrt[2]{|ui1 - uj1|^2 + |ui2 - uj2|^2 + \cdots + |uik - ujk|^2} \quad \text{–(11)}$$

3.Manhattan Distance:
q=1

$$S(i,j) = |ui1 - uj1| + |ui2 - uj2| + \cdots + |uik - ujk| \quad \text{–(12)}$$

Euclidean approach is the most generally utilized technique to compute the distance test sample and trained data values.

3 .2.3How to choose a K value
K indicates the parameter that is the number of the nearest neighbors. Tracking down the best k worth to accomplish the greatest precision of the model is a difficult errand. There is no pre-characterized measurable strategy to recognize the k worth to accomplish amazing precision. The only method to find k value which attain impressive accuracy is to use Brute force method which means we need to find accuracy for different k value. The K values of neighbors 1 to 20 and the neighbor that gives the highest accuracy is considered for the prediction.

3 .3Support Vector Machine (SVM)
SVM is a supervised technique. This algorithm is utilized for both classification and regression studies. In this algorithm, data is plotted in n-dimensional space(uses coordinates).SVM can be classified into linear and nonlinear types. In our work, we use data that is linearly separable so we use the linear SVM classifier.
The classification of classes is done by finding the optimal hyperplane, hyperplanes are the boundaries that divide the classes into categories. The line is a hyperplane in two-dimensional space. The line is sufficient in two-dimensional space to separate the classes.
For example, consider the equation S0+(S1*U1)+(S2*U2)=0
B0, B1 are coefficients and B2 is the intercept of the line.K1, K2 are input points. This line is used for the classification. Above the line, the value returned by the equation is more than zero and the data value belongs to the category "0".Below the line, the value returned by the equation is less than zero and the data point belongs to the category "1".A point that yields a value close to zero is difficult to classify. Margin is referred to as the distance between the closest data point and the line.The ideal line can isolate the classes on the off chance that it has the biggest edge. This line is known

as maximal margin hyperplane. This margin is registered by utilizing the perpendicular distance between the nearest highlight the line and the line. These points are crucial for describing the line and for the classifier construction and the data values are called support vectors. Hyperplanes are supported and defined by support vectors.

3 .4 Dataset Used

3 .4.1Heart Dataset

For predicting the occurrence of heart diseases we have used the "Heart Disease Dataset" by UCI. This dataset consists of 13 medical predictor features and one target feature. The attributes are as follows chol, cp, trestbps, age, fbs, sex, restecg, exang, slope, thal, ca, oldpeak, thalach. The dataset contains 303 instances and 75 attributes.

3 .4.2Diabetes Dataset

For predicting the occurrence of Diabetics diseases we have used the "Pima Indians Diabetes Dataset" by Kaggle. This dataset contains 8 medical predictor features and one target feature. The attributes are as follows Blood pressure, Pregnancies, Glucose, SkinThickness, BMI, Insulin, Age, and DiabetesPredigree.

3 .4.3Breast Cancer Dataset

For predicting the occurrence of breast cancer diseases we have used the "Breast Cancer Wisconsin (Diagnostic) Data Set" by Kaggle. This dataset contains 31 medical predictor features and one target feature. Some of the important attributes are as follows diagnosis, id, radius-mean, texture-mean, perimeter-mean, area-mean, smoothness-mean, compact-mean, concavity-mean, and concavity points-mean.

4 . Results

The actions performed in this work are done by the Laptop with an i5 processor and developed the code using python. The algorithms used in this work are Logistic Regression, KNN, SVM, and the accuracies are calculated using the cross-validation with factor cv as 10. The accuracies of each disease are illustrated through bar graphs. The datasets of diseases are divided into training and test datasets for classification.

AUTHORS	TECHNIQUE USED	DISEASE	ACCU RACY
M. Marimuthu, S.Deivarani, R.Gayathri.	KNN	Heart disease	83.60%
Purushottama, Richa Sharma, Dr. Kanak Saxena.	SVM	Heart disease	70.59%
K. Ramesh,G.Subbalakshmi and M. Chinna Rao.	Naïve Bayes	Heart disease	52.33%
V.Rajesh, A.Vamsi Akki Reddy,Naveen Kishore G K.Sumedh, T.Rajesh Sai Reddy.	Logistic Regression	Diabetes	72.39%
Gavin Pinto, Sunil Jangid, Radhika Desai.	SVM	Diabetes	64.92%
Deepti Sisodia, Dilip Singh Sisoda.	Naive Bayes	Diabetes	76.30%
Dr.J.Ajayan, Dr.B.Santhosh Kumar, T.Daniya,	KNN	Breast Cancer	83.33%
Shaik Subhani, K. Pravalika, Ch. Shravya,	Logistic Regression	Breast Cancer	92.10%
Mümine KAYA KELEŞ	Random Forest	Breast Cancer	92.20%

Table1: Comparison Table

The accuracies mentioned in the table1 have been surpassed by our work. The highest accuracies achieved by our work for Diabetes, Heart, and Breast cancer are as follows 77.60%, 83.84%, and 94.55% respectively using Logistic regression for diabetes and Breast cancer and KNN for Heart disease.

Disease	Logistic Regression	SVM	KNN	Best Accuracy
Heart Disease	82.50	78.87	83.84	83.84
Diabetes	77.60	75.64	75.52	77.60
Breast Cancer	94.55	91.38	92.55	94.55

Table 2: Accuracies of different algorithms.

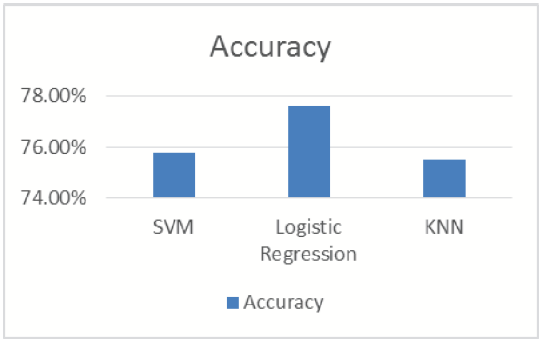

Figure2: Represent the accuracy values for diabetes.

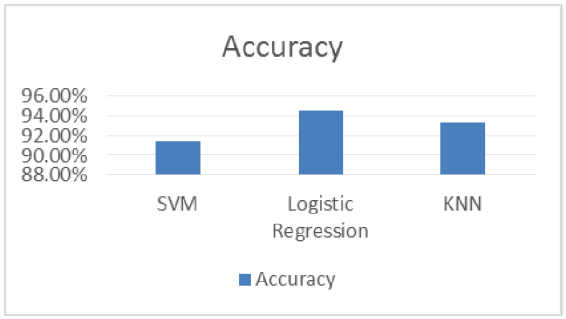

Figure3: Represent the accuracy values for Breast cancer

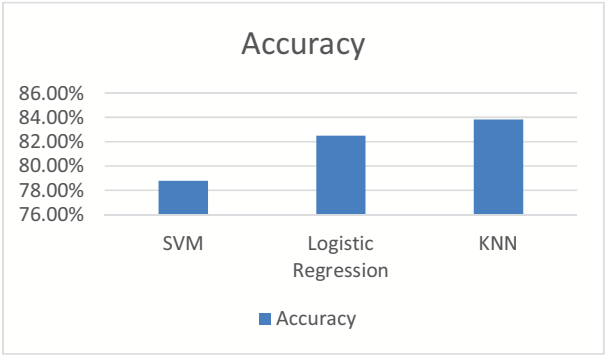

Figure4: Represent the accuracy values for Heart Disease.

5. Conclusion

The proposed work brings diabetes, heart disease, and breast cancer under a single platform by deploying the trained models using the flask API framework which is a lightweight framework.

Three classification algorithms are used for training the models, in which the logistic regression gave

good accuracy values for the disease prediction of diabetes and breast cancer and KNN for the disease prediction of heart disease. KNN's highest accuracy is calculated by picking the highest value obtained from 1 to 21 neighbors. In the future, we can expand this work by adding more diseases that are trained by machine learning models and also can include the disease that involves deep learning models.

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.