Heart Disease Diagnosis using SVM and Logistic Regression Model of Machine Learning Project Report

Aman Raj (IIT2018038), Chandramani (IIT2018041), Hrutvik Nagrale (IIT2018088),

Ravi Kumar (IIT2018094), Vaibhav Shukla (IIT2018098), Anshul Ahirwar (IIT2018099)

Guide : GC Nandi

V Semester B.Tech Information Technology,

Indian Institute of Information Technology Allahabad, Prayagraj

Abstract: Machine learning tools are widely used in various fields of science and technology. They gave meaningful and classified information. As the number of deaths in humans from sudden heart failure due to various medical reasons is increasing, a prediction system that helps the patient to be vigilant about his condition is needed. Our project objective is to detect whether patients have heart disease or not by using the given number of features from patients. The motivation of our project is to save human resources in medical centers and improve accuracy of diagnosis. In our project we have used two methods to detect heart disease such as Logistic Regression and SVM.

Keywords: Heart Disease, Logistic Regression, Support vector machine (SVM), Confusion Matrix.

I. INTRODUCTION

Cardiovascular disease, also known as heart disease, is a widespread and persistent problem in the field of medical analysis. The burden of cardiovascular diseases has increased rapidly worldwide in recent years. Although these diseases have been shown to be the leading cause of death, they have been considered the most manageable and preventable disease. Most often, the blockage of the arteries causes a heart attack. This happens when the heart is not efficiently pumping blood around the body. High blood pressure is also a major cause of heart disease. Likewise, there are many other reasons for contracting heart disease, including obesity, poor diet, high cholesterol, and lack of physical activity. So prevention is very necessary. Knowing about heart disease is important for prevention. About 47% of people die outside of the hospital and this shows that they don't act on the first warning signs.

Detection is a major challenge in heart disease. It is difficult to predict whether or not a person has heart disease. There are tools that can predict heart disease, but they are expensive or inefficient in calculating the likelihood of heart disease in humans. In the case of India, access to good doctors and hospitals in rural areas is very limited. Heart disease is a major challenge in medical science. Machine learning could be a great option for predicting heart disease in humans.

We want to predict whether the patient has heart disease or not using given features of the patients. If such a prediction is accurate enough, we can not only avoid diagnostic errors but also save personnel. If a patient without heart disease is diagnosed with heart disease, they panic needlessly, and if a patient with heart disease is not diagnosed with heart disease, they lose the best chance of curing their disease. This misdiagnosis is painful for both patients and hospitals. With accurate predictions, we can solve unnecessary problems.

The input for our algorithm consists of 14 features with numerical values. We use two algorithms like Logistic Regression and SVM to generate a binary number 1 or 0. 1 indicates that the patient has heart

disease and 0 indicates that the patient does not have heart disease.

II. MACHINE LEARNING

Machine learning is widespread in almost many areas of the world, including the healthcare industry. Machine learning is an application of artificial intelligence (AI) that enables systems to automatically learn from experience and improve without being explicitly programmed. Machine learning in its most basic form is also about using algorithms to analyze data, learn from it, and then make a determination or prediction about something in the world. There are two main categories of problems that machine learning often solves, namely regression and classification. Regression algorithms are primarily used on numeric data, and classification problems include binary problems and multi-category problems. Machine learning algorithms fall into two categories, Supervised learning and Unsupervised learning. In principle, supervised learning takes place using previous knowledge in the output values, while unsupervised learning does not have any predefined names. Hence the purpose is to infer the natural structures in the dataset given. Therefore, the choice of machine learning algorithm should be carefully evaluated.

III. DATASET AND FEATURES

Our dataset is based on UCI heart disease data set, particularly taken from the Cleveland database. This dataset contains 76 attributes but we have taken a subset of 14 of them in order to reduce the noise. The features are explained below:

- age: age in year
- sex: either male or female
    - Value 0: female
    - Value 1:male
- cp: chest pain type
    - Value 1: typical angina
    - Value 2: atypical angina
    - Value 3: non-anginal pain
    - Value 4: asymptomatic
- trestbps: resting blood pressure (in mmHg)
- chol: serum cholesterol (in mg/dl)
- fbs: fasting blood sugar
    - Value 0: fasting blood sugar < 120 mg/dl
    - Value 1: fasting blood sugar > 120 mg/dl
- restecg: resting electrocardiographic results
    - Value 0: normal
    - Value 1: having ST-T wave abnormality (ST and/or T wave inversion)
    - Value 2: showing probable or definite left ventricular hypertrophy
- thalach: maximum heart rate achieved
- exang: exercise included angina
    - Value 0: No
    - Value 1: Yes
- oldpeak: ST depression included by exercise relative to rest
- slope: the slope of the peak exercise ST segment
    - Value 1: upsloping
    - Value 2: flat
    - Value 3: downsloping
- ca: number of major vessels (0-3) colored by fluoroscopy
- thal
    - Value 3: normal
    - Value 6: fixed
    - Value 7: reversible
- num: angiographic disease status
    - Value 0: 50% diameter narrowing, no heart disease
    - Value 1: 50% diameter narrowing, heart disease

We split the dataset into 8:2 ratio for training and testing respectively resulting in 242 instances for training and 61 for testing. To avoid overfitting the dataset was normalised. In this project we have considered 14 attributes to train models for SVM and logistic regression in order to predict (attribute: num) if a patient has been suffering from a heart disease or not.

IV. METHODS

In this project, we have implemented two models, Logistic Regression model and SVM model for prediction of heart disease.

A. Logistic regression :

The logistic regression is a supervised learning technique that can be used to classify data into binary or multiple classes. Here, we have used the binary model in our heart disease prediction project to predict

if a person has a heart disease or not, given various parameters. In our model, we've used the sigmoid function below:

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

This sigmoid function maps any real value between 0 and 1. We can say that this value is our probability of a person being diagnosed with heart disease. We've assumed the threshold value to be 0.5. If the resultant probability is greater than 0.5, we will say that the person has a heart disease otherwise no heart disease.

B. SVM (Support Vector Machine) :

SVM are supervised learning models used for classification and regression. It takes a set of input data, particularly features and predicts for each input, which two possible classes forms the output. It aims to find a hyperplane in a dimension that classifies the dataset.

The equation of the hyperplane is of the form,

$$w^T x + b = 0$$

w : weight vector
x : input vector
b : bias

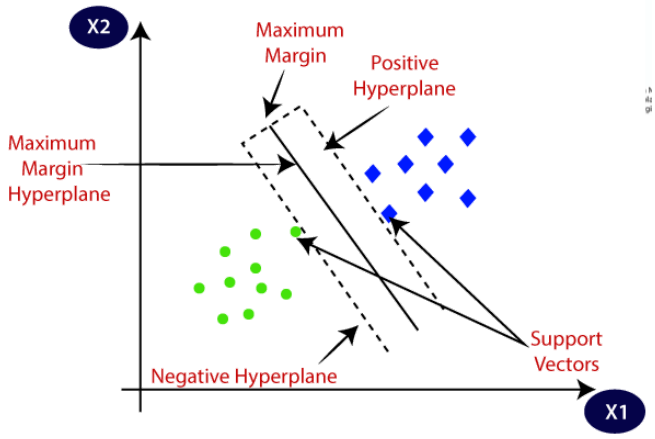


Fig 1: Classification by SVM

The objective is to maximize the margin separating the two classes and it is given by

$$\frac{w^T(x_+ - x_-)}{||w||} = \frac{2}{||w||}$$

Here $w^T x_+ + b = 1$ and $w^T x_+ + b = -1$. Maximizing margin requires $\frac{2}{||w||}$ to be maximised, which is equivalent to minimizing $||w||^2$. We find the vector w which minimizes the cost function using Gradient descent.

$$||w||^2 + C\sum_{i}^{N} max(0, 1 - y_i f(x_i))$$

We have implemented both soft margin and hard margin SVM with different kernels. Soft margin SVM just negates the effect of outliers which may refrain formation of a hyperplane (by hard margin SVM) of dataset which may not be linearly separable.

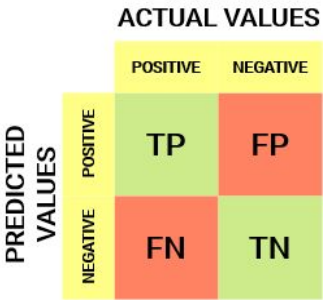
V. EXPERIMENTS / RESULTS

The design of the simulator is highly dependent on the level of implementation implemented.

Confusion Matrix Outcomes :

This is used to show the summary of the prediction results, including correct and incorrect results in a classification problem. This is used not only for errors but also for error types.

For a binary classification we have,



True Positives (TP): Predicted cases (with the disease).

False Positives (FP): Cases believed to be true but actually not suffering from the disease.

False Negatives (FN): Cases where it is predicted that you will not have the disease, but actually have the disease.

True Negatives (TN): Cases in which this is unlikely to be the case and they do not have the disease.

A. Logistic regression :

The Loss Function which we used for the logistic regression is given below:

$$L(\hat{y}, y) = -(y log(\hat{y}) + (1-y) log(1-\hat{y}))$$

where $\hat{y}$ is the predicted value and y is the actual value.
The results for *learning rate = 0.001* are:
Training set accuracy: ***85.84905660377359 %***
Test set accuracy: ***81.31868131868131 %***

Learning rate	Training set accuracy	Test set accuracy
0.00001	84.43	81.31
0.00005	85.84	81.31
0.0001	85.84	81.31
0.0005	87.26	80.21
0.001	86.79	81.31
0.005	86.79	79.12
0.01	86.79	79.12
0.05	86.79	79.12
0.1	86.79	79.12
0.5	86.79	79.12

The following graph illustrates the *loss vs number of iterations* for different values of *learning rate.*

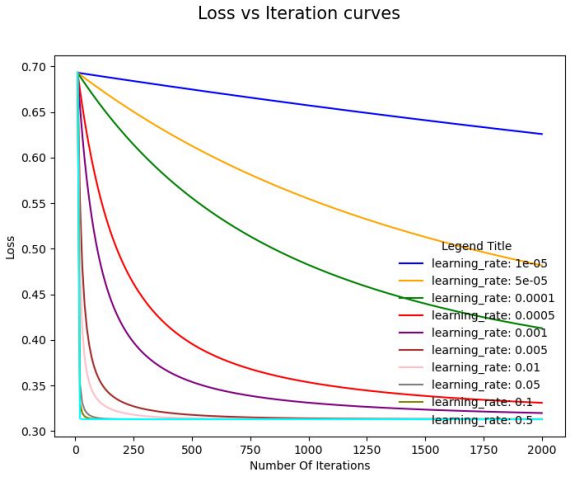


Fig 2: *loss vs number of iterations* for different values of *learning rate.*
Following is the confusion matrix for Logistic Regression on the test set.

43	13
4	31

Following is the confusion matrix for Logistic Regression on the train set.

103	16
14	79

B. SVM (Support Vector Machine) :

a) Soft margin SVM

In soft margin SVM we try to minimize the outliers by adding cost to outliers in the cost function and require the solution of following optimization problem:

$$\text{Min}_{(w,\ \xi\ ,b)}\ \frac{1}{2} w^T w + C\ \Sigma^{n}_{i=1}\ \xi_i$$

$$\text{subject to}\quad Y^{(i)}(w^T s^{(i)} + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0\ ,\ \text{for i=1,2,....n.}$$

Training set accuracy: ***87.60330578512396 %***
Test set accuracy: ***81.9672131147541 %***

Following is the confusion matrix for soft margin SVM on the test set.

27	5
6	23

Following is the confusion matrix for soft margin SVM on the train set.

117	16
14	95

b) SVM with polynomial kernel

Kernel functions are used to map data points to higher dimensional space and then SVM finds a linear separating hyperplane with maximal margin in this higher dimensional space. In polynomial kernel the kernel function which is used to map data to hughes dimension is :

$$K(x_i, x_j) = (\gamma\, x_i^T\, x_j + r)^d,\ \gamma > 0.$$

Training set accuracy: ***100%***
Test set accuracy: ***77.04918032786885 %***

Following is the confusion matrix for SVM with polynomial kernel on the test set.

30	11
3	17

Following is the confusion matrix for SVM with polynomial kernel on the train set.

131	0
0	111

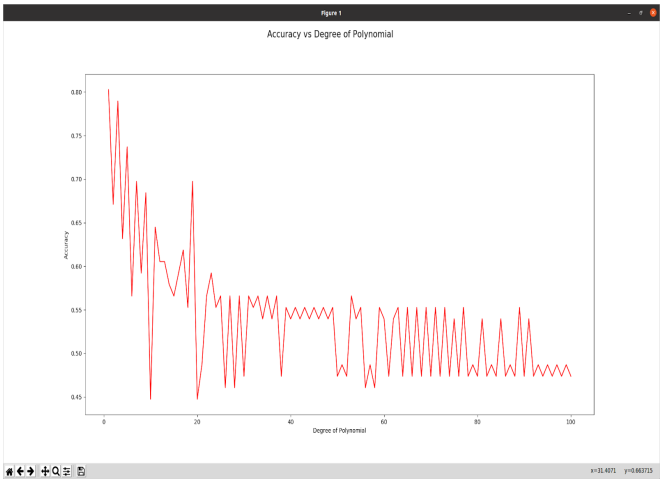


Fig 3: *Accuracy vs Degree of Polynomial.*

The accuracy decreases for higher degree polynomial because it overfits the training data and performs poorly on the test data.

c) SVM with RBF kernel

A RBF kernel, like a polynomial kernel, non-linearly maps the data to higher dimensional space. It is preferred over polynomial kernels because it has less hyperparameters and hence the complexity is less in comparison to polynomial kernels. The RBF kernel function is :

$$K(x_i, x_j) = \exp(-\ \gamma\ |xi\ - xj|^2)\ ,\ \gamma > 0.$$

Train Accuracy: **97.93388429752066%**
Test Accuracy: **83.60655737704919%**

Following is the confusion matrix for SVM with Gaussian kernel on the test set.

30	7
3	21

Following is the confusion matrix for SVM with Gaussian kernel on the train set.

130	4
1	107

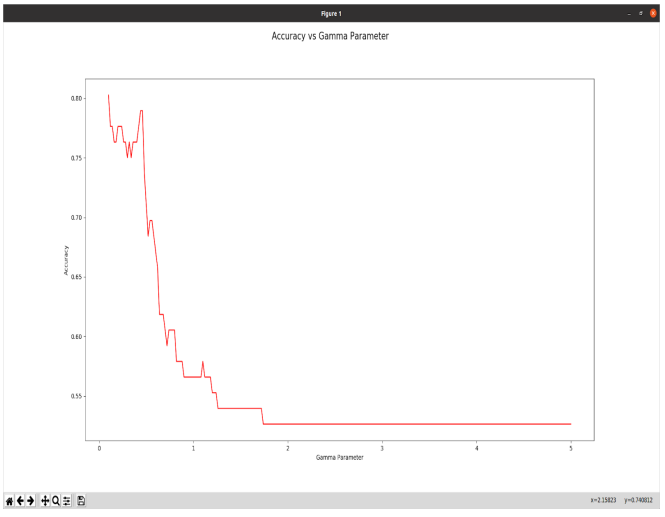


Fig 4: *Accuracy vs Gamma Parameter.*

Gamma parameter is the inverse of standard deviation of RBF kernel. Intuitively a small gamma value will define gaussian function with a large variance and vice versa. Initially increasing the gamma value makes the

appropriate generalization but when accuracy starts to decrease, the SVM starts to overfit the training data.

d) SVM with linear kernel

A linear kernel unlike other mentioned kernels maps the data linearly in higher dimensional space and hence it cannot capture the nonlinear relation between class labels and attributes. The linear kernel function is :

$$K(x_i, x_j) = x_i^T x_j.$$

Training set accuracy:***87.19008264462809%***
Test set accuracy: ***81.9672131147541 %***

Following is the confusion matrix forSVM with linear kernel on the test set.

27	**5**
6	**23**

Following is the confusion matrix for SVM with linear kernel on the train set.

118	**18**
13	**93**

TABLE 1. RESULTS OF DIFFERENT METHODS

Methods	***Train accuracy***	***Test accuracy***
Logistic Regression	*85.85 %*	*81.32 %*
SVM Soft Margin	*87.60 %*	*81.97%*
SVM with Gaussian Kernel	*84.71%*	*81.97%*
SVM with Linear Kernel	*87.19%*	*81.97%*
SVM with Polynomial Kernel	*100.0%*	*77.05%*

VI. CONCLUSION

We have used some libraries ex scikit-learn (only to split the dataset into train and test sets), numpy, pandas, matplotlib and cvxopt to implement this project. After the experiments, we have found that the SVM Soft Margin gives us the overall best accuracy.

REFERENCES

[1]https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data

[2]https://drive.google.com/file/d/1eiIJpWlRbTO-qaEHupj4gIFXXYbYh4dJ/view?usp=sharing - OUR PROJECT DATASET.