ABSTRACT

The amount of people suffering from heart diseases is immense throughout the world let alone in India. To address this we can monitor the patients throughout their lives but that isn't enough cause we cannot predict the problem at hand with ease. Certain aspects of the human heart and body when compiled together give us a basis for a study of if that prediction could save lives. Today there are many models which aim to do so in different ways. We have come up with a few simplistic models and have done an in-depth analysis on how we can better the prediction algorithm. Research has attempted to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk using homogenous data mining techniques. We have used Logistic regression and support vector machine algorithms to get accurate results on a trained model with a dataset of 303 entries. In this paper we further explain how these models are beneficial and how the application of linear algebra in these models makes it more apt for the problem.

Keywords: Heart Disease, Data Mining, Logistic Regression, Support Vector Machine, Radial Basis Function, Linear Algebra

INTRODUCTION

In India, out of the estimated population of more than 1.27 billion dispersed across various geographical regions, about 45 million people suffer from coronary artery disease. In the medical industry huge amounts of data is available but people are not able to extract the important information about the factors that cause cardiovascular disease. A lot of lives could be spared by diagnosing on schedule. Along these lines, diagnosing the syndrome is significant and an exceptionally muddled undertaking. There is a shortage of physicians, especially in the rural areas. There is a lack of awareness among the people regarding healthy lifestyle habits. Because of all these factors, people do not visit the doctor/hospital for the regular medical checkups and the disease is not diagnosed in time. The disease is usually detected when the symptoms like weakness of the physical body, swollen feet, shortness of breath, cold sweats and fatigue develop. Thus, there is a great necessity to develop cost effective and easily accessible tools which can diagnose the heart diseases in time. Research using data mining models have been applied to diseases such as diabetes, asthma, cardiovascular diseases, AIDS, etc. Research in the field of cardiovascular diseases using data mining techniques has been an ongoing effort involving prediction, treatment and risk score analysis with high levels of accuracy. Data mining will assist doctors to carry out their diagnosis. It will help doctors to emphasize on some informative knowledge to predict the disease more quickly. Various techniques of data mining such as naïve Bayesian classification, artificial neural networks, support vector machines, decision trees, logistic regression, etc. have been used to develop models in healthcare research.

Tasks in data mining are split into two type's i.e. predictive and descriptive tasks. Predicting the value of an individual attribute on the basis of another attribute is done in Predictive tasks. Predictive tasks include classification technique. Descriptive tasks summarize the relationship between data and it determines patterns. Descriptive tasks include clustering techniques. Data preprocessing is an

essential step in the data mining process. The performance of such a prediction system not only depends on the choice of machine learning algorithm used but also on the quality of raw medical data. Data-gathering methods are generally loosely controlled, occurring in out-of-scope values, impractical combinations of data, missing values, etc. Evaluating data that has not been carefully hidden for such problems can produce inaccurate results. Thus, the representation and data quality is first and most important before running an analysis. If there is much inappropriate and redundant data present or noisy and unreliable data, then knowledge discovery at the time of the training phase is also challenging. Data preprocessing tasks include data cleaning, data integration, data transformation, data reduction and data discretization.

The methodology aims to accomplish two goals: the first is to primarily present a predictive framework for heart disease, and the second is to compare the efficiency of merging the outcomes of multiple models as opposed to using a single model.

LITERATURE SURVEY

Prediction of heart disease has been focused by several studies. Different data mining tools and applied various data mining techniques are used for diagnosis and achieving various conclusions. The goal of all is to achieve better accuracy and to make the system more efficient so that it can predict the risk of heart attack. Patients with high risk of having heart attack can be easily identified by the professionals with the help of health care associated with risk factor knowledge.[1]

Coronary heart disease has various precautionary measures like no smoking, exercise regularly, maintaining blood cholesterol and weight of the body, properly maintaining diabetes and high blood pressure levels. Early detection is needed for the people who are at very high cardiovascular risk or people with cardiovascular disease. Hence, more efficient methods of cardiovascular disease are of great concern.[2]

Comparing the accuracies across multiple data sets with different parameters arrives at different results which do not provide a just basis for comparison. One of the bases on which the papers differ are the selection of parameters on which the methods have been applied. Many authors have specified different parameters and databases for testing the accuracies. Xing et al[7], conducted a survey of 1000 patients, the results of which showed SVM to have 92.1% accuracy, artificial neural networks to have 91.0% and decision trees with 89.6% using TNF, IL6, IL8, HICRP, MPO1, TNI2, sex, age, smoke, hypertension, diabetes, and survival as the parameters. Similarly, Chen et al[8], compared the accuracy of SVM, neural networks, Bayesian classification, decision tree and logistic regression. Considering 102 cases, SVM had the highest accuracy of 90.5%, neural networks 88.9%, Bayesian 82.2%, decision tree 77.9%, and logistic regression 73.9%. [1]

PROPOSED FRAMEWORK

I. DATA SET

Understanding the dataset is the most crucial step of the complete framework. Making statements and understanding relatability of each component in the dataset will give meaning to this project in different ways. Our research is based on the data vectors:

Attribute	Description
Age	how many years old Range[25 -110]
Gender	having value 1 for 'Male' and value 0 for 'Female'
Chest pain	value 1 for 'typical angina' value 2 for 'atypical angina' value 3 ' non-angina' value 4 'asymptomatic'
Resting Blood Pressure	blood pressure mm Hg range [60 - 200]
Cholesterol	cholesterol measured in mm/dl range[120 -600]
FBS (Fasting Blood Sugar)	value 1 for '>120 mg/dl' value 0 for '<=120 mg/dl
Resting ECG (Electro Cardio Graph	value 0 for 'normal' value 1 for 'abnormal' value 2 for 'showing probable'
Maximum Heart rate	heart count per minute range[60-200]
Exercise Induced Angina	value 0 for NO value 1 for YES
Old peak	ST curve relative to Rest range[0-6]
Slope	The slope of peak exercise ST segment value 1 for 'upsloping' value 2 for 'flat' value 3 for 'down sloping'
Coronary Artery	count of major vessels colored by floursospy range[0-3]
Thalassemia	Bone marrow expand relative congestive heart failure value 3 for 'normal' Value 6 'fixed defect' value 7 'reversible defect'
Class Label	0 for 'absence of heart disease' 1 for 'presence of heart disease'

II. DATA PREPROCESSING

Data does not serve its purpose until it is processed and understood by the machine. The step involves transforming the data, which involves removal of missing fields, normalization of data, and removal of outliers. These steps define our dataset and make it easy for us to get a hold of which fields are more dependent for the result and which ones are less. Appropriate graphs for the preceding help us get to know the data better and clearly analyse it for further training of the data. For SVM, data points were automatically centered at their mean and scaled to have unit standard deviation. No changes need be made to the data sets for logistic regression.

III. METHODOLOGY

Logistic Regression

Logistic regression is a machine learning algorithm used for classification. It is based on the concept of probability. Logistic regression is used to assign observations to a discrete class. Transforming output is done using the sigmoid logic function. The logistic regression hypothesis tends to limit the cost function in range between 0 and 1. Therefore, linear functions can not represent as it can have a value >1 or <=0, which is not possible according to the regression hypothesis.

A minimization function is used that is the cost function. It uses the Log Loss i.e. the logarithmic loss which measures the performance of the model where the prediction input value is the probability between the zero and one. The Log loss is the uncertainty of the prediction which is based on how much it varies from the actual label. Cost function which helps the learner to correct or change the behavior to minimize the mistakes is :

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} y^{(i)} \log\left(h_\theta(x^{(i)})\right) - \left(1 - y^{(i)}\right)\log\left(1 - h_\theta(x^{(i)})\right) + \frac{\lambda}{2m}\sum_{i=1}^{n} \theta_j^2$$

m = number of instances n = number of attributes y = class label
x = train data features
θ = coefficients λ =learning rate

lambda is also called the regularization factor.

The last term in the cost function corresponds to the Regularization term.

Regularization : It is used to overcome the problem of overfitting which occurs when there are too many features involved. Due to overfitting the model fails to generalise to new inputs though it does for the training data. The few methods which address the problem of Overfitting are:

1) Reduce the number of features :

- manually select which features to keep.
- this is called a model selection algorithm.

2)Use Regularization

- keep all the features , but reduce the magnitude of parameters Θj .
- this method works well when we have lot of features, each of which contributes to hypothesis.

Gradient descent :It is an optimization method which is used to find the parameters or the coefficient of the cost function. Gradient descent is a repeated process in order to get the coefficients to minimize the cost function. The Gradient descent is calculated for both the classes to get the pair of a coefficient for both class labels. The goal here is to continue the procedure to try the different values for the coefficients, evaluating their cost and selecting the new coefficient that is having the slightly lower cost. Considering this coefficient and storing them in the model. Gradient descent is calculated as the following:

$$\theta_{j:} = \theta_j - \frac{1}{m}\sum_{i=1}^{m} (h_\theta(x^i) - (y^i))\, x_j^i + \frac{\lambda}{m}\theta_j$$

m = number of instances x = train data features
y = class label θ = coefficients λ = learning rate

Sigmoid function: is used as the hypothesis function for logistic regression . This takes the real input values and output values between the 0 and 1 for logistic function. This is interpreted as taking log odds and having the output probability. Generally sigmoid function is used to map predictions to probability it is defined as:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

x = test data features θ = coefficients

The graph of the hypothesis function is shown below:

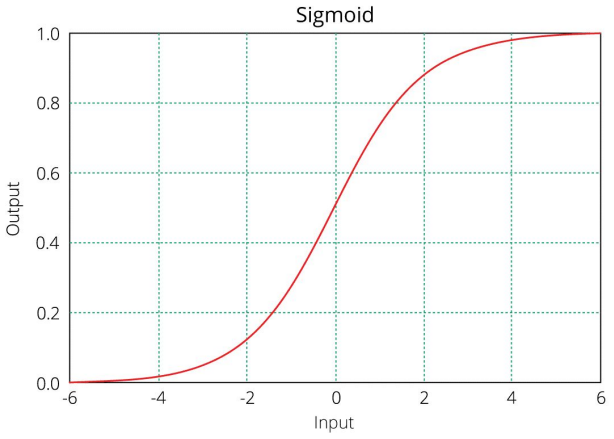

Figure 4: Sigmoid function

Whenever a test data is passed it calculates the value based on the parameters stored in the model. It calculates the probability of each class label. We return the maximum probability value of the class label Xi.

The test data contains the thirteen attributes that we need to pass and calculate for both the classes it will return the two values we take the maximum value of

two values we will return the class label which is having the maximum probability.

SVM (Support Vector Machines)

SVMs are a set of related supervised machine learning methods used for classification and regression, they belong to the family of generalized linear classification.The Support Vector Machine was first proposed by Vapnik and has since attracted a high degree of interest in the machine learning research community. The speciality of SVM's is that they minimize the classification error and maximize the geometric margin. The data points are isolated in such a way that a straight line could be drawn to separate them into classes. This shows that the data is linearly separable , this line is called separating hyperplane and the data points close to this line are called support vectors. SVM's can also fit nonlinear complex models by using different kernels which are described in the following paragraph**.** They are developed by improving logistic regression hypothesis, Several recent studies have reported that the SVM generally is capable of delivering higher performance in terms of classification accuracy than the other algorithms present.

In machine learning, the kernel is a technique that is used to solve the nonlinear problem with the use of linear classifiers and involved in exchanging linearly non-separable data into linearly separable data. The idea behind this concept is linearly non-separated data in N-dimensional space might be linearly separate in high M-dimensional space.

There are several kernels functions some of them listed below here :

Polynomial Type: is well known for nonlinear modeling and is represented as

K (a, b) = (a, b) d

Gaussian Radial Basis Type: Radial basis functions mostly with Gaussian form and represented by

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

Exponential Radial basis: function produces a bitwise linear solution that will be useful when discontinuities are satisfactory.

In addition to them, there are many more functions such as multi-layer perceptron, Fourier, additive, and tensor products type.

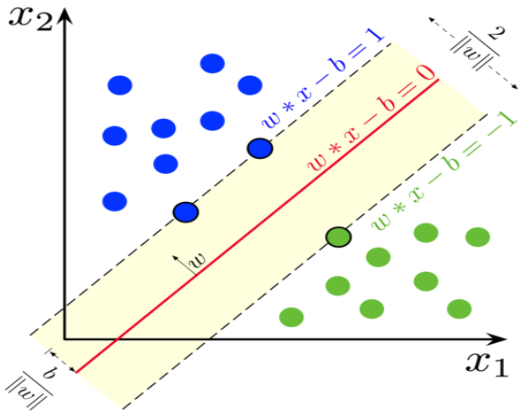

Figure 1: maximum-margin hyperplane for an SVM trained with samples of two classes.

w: represents the unit normal vector to the support vectors .

x: represents the data vectors. b is a constant.

Cost Function

$$J(\theta) = C[\sum_{i=1}^{m} y^{(i)} Cost_1(\theta^T(x^{(i)}) + (1 - y^{(i)}) Cost_0(\theta^T(x^{(i)})] + \frac{1}{2}\sum_{j=1}^{n} \theta_j^2$$

m = *number of samples*, n = *number of features*

The general intuition of the cost function is shown above. The last term of cost function is for

regularization which is described in Logistic regression. The cost function is used to train the SVM. By minimizing the value of J(theta), we can ensure that the SVM is as accurate as possible. In the equation, the functions cost1 and cost0 refer to the cost for an example where y=1 and the cost for an example where y=0. For SVMs, cost is determined by kernel (similarity) functions.

Choosing the kernel comes in handy when we have multiple choices. The Radial basis function kernel, also called the RBF kernel, or Gaussian kernel, is a kernel that is in the form of a radial basis function. The gaussian radial basis type kernel is a stationary kernel, which means that it is invariant to translation. Also, it is isotropic in behaviour i.e. the scaling by γ occurs in the same amount in all directions. Another property of RBFs is that it is infinitely smooth. The RBF kernel is defined as KRBF(x, x 0) =

$$k(\mathbf{x_i}, \mathbf{x_j}) = \exp(-\gamma \|\mathbf{x_i} - \mathbf{x_j}\|^2)$$

where γ is a parameter that sets the "spread" of the kernel

Kernel expresses a measure of similarity between vectors in the case of RBF kernel, it represents this similarity as a decaying function of the distance between the vectors (i.e. the squared-norm of their distance). That is, if the two vectors are close together then, ||x − x`|| will be small. Then, so long as γ > 0, it follows that -ϒ||x − x`||^2 will be larger. Thus, closer vectors have a larger RBF kernel value than farther vectors. This function is of the form of a bell-shaped curve.

The γ parameter sets the width of the bell-shaped curve. The larger the value of γ the narrower will be the bell. Small values of γ yield wide bells. This is illustrated in the figure below.

Figure 2: (a) Large □ (b) Small □

In this model the concept of vector inner product from linear algebra is used for deriving the large margin classifier i.e the hyperplane which acts as a binary classifier for the proposed dataset.

RESULTS

Training accuracy for logistic regression came out to be 86.79% and that for SVM came out to be 93.40%. Further when we tested these models for the unexplored dataset the accuracies were 86.81% and 87.91% for logistic regression and SVM respectively. The dataset was divided into 0.7 for training purposes and the remaining was used for testing purposes.

	Model	Training Accuracy %	Testing Accuracy %
0	Logistic Regression	86.79	86.81
1	Support Vector Machine	93.40	87.91

PERFORMANCE ANALYSIS

Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the actual values are

known.

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Figure 3: Confusion matrix

True positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.

True negatives (TN): We predicted no, and they don't have the disease.

False positives (FP): We predicted yes, but they don't actually have the disease.

False negatives (FN): We predicted no, but they actually do have the disease.

Accuracy:It estimates Overall, how often is the classifier correct. It is represented as :

(TP+TN) / (TP+TN+FP+FN)

Precision: It is about when the model predicts how often it is correct. It is represented as :

TP / (TP+FP)

Recall: It is the fraction of the total amount of relevant instances that were actually retrieved. It is represented as :

TP / (FN+TP)

F1 Score: It considers both the precision p and the recall r of the test to compute the score. It is represented by :

2 x ((precision x recall) / (precision+recall))

for **Logistic Regression**

```
Test Result:
================================================
Accuracy Score: 86.81%

_____________________________________________________
Classification Report:  Precision Score: 86.54%
                        Recall Score: 90.00%
                        F1 score: 88.24%

_____________________________________________________
Confusion Matrix:
 [[34  7]
 [ 5 45]]
```

for **Support Vector Machine**

```
Test Result:
================================================
Accuracy Score: 87.91%

_____________________________________________________
Classification Report:  Precision Score: 89.80%
                        Recall Score: 88.00%
                        F1 score: 88.89%

_____________________________________________________
Confusion Matrix:
 [[36  5]
 [ 6 44]]
```

CONCLUSION

This paper proposes a framework using combinations of support vector machines and logistic regression to arrive at an accurate prediction of heart disease. Major life threatening disease that leads to death is Heart disease. Heart is the most essential organ of the human body as life is dependent on proficient working of heart. The consultant of doctor's determination can make without the advice of specialists because of the software developed by the advancement in computer technology. Linear Algebra will play a vital role in medical areas especially use for prediction. The framework can also be extended for use on other models such as neural networks, ensemble algorithms etc.

REFERENCES

[1] A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL) Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu

[2] Prediction of Cardiovascular Diseases using Support Vector Machine and Bayesian Classification Prashasti Kanikar Assistant Professor (Computer Engg.) Disha Rajeshkumar Shah M Tech (Computer Engg.)

[3] Determination of Significant Features for Building an Efficient Heart Disease Prediction System Ekta Maini, Bondu Venkateswarlu, Arbind Gupta

[4] A Cardiovascular Disease Prediction using Machine Learning Algorithms Rubini P E, Deeksha G S, B Varshaa Shree, Deepa N, Abhinav Srivastava

[5] Support vector machine the most fruitful algorithm for prognosticating heart disorder M. Murugesan , R. Elankeerthana

[6] Heart Disease UCI at https://www.kaggle.com/ronitf/heart-disease-uci

[7] Yanwei Xing, Jie Wang and Zhihong Zhao Yonghong Gao 2007 “Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease” Convergence Information Technology, 2007. International Conference November 2007, pp 868-872.

[8] Jianxin Chen, Guangcheng Xi, Yanwei Xing, Jing Chen, and Jie Wang 2007 “Predicting Syndrome by NEI Specifications: A Comparison of Five Data Mining Algorithms in Coronary Heart Disease” Life System Modeling and Simulation Lecture Notes in Computer Science, pp 129-135.

[9] Peter Harrington, “Machine Learning in Actions”, Published in April 16th 2012 by Manning Publications.

[10] Yuxi Liu“Python Machine Learning by example”, Published by Packt Publishing 2019.

[11]https://data-flair.training/blogs/svm-kernel-functions/ for kernels.

[12] https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/ for performance Analysis.

[13]https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/ regarding confusion matrix.