INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 04 ISSUE: 07 | JULY -2020 ISSN: 2582-3930

Prediction of heart disease and diabetes using machine learning

1.Akarsh Singh*, 2. Akshat Agrawal*, 3. Ayush Bhargava*, 4. Srijan Yadav*, 5 H K Vedamurthy**

*Dept of CSE, Siddaganga Institute of Technology, Tumakuru
** Assistant Professor, Dept of CSE, Siddaganga Institute of Technology, Tumakuru

Abstract - Classifying data is oneofthe most well-known tasks in Machinelearning. Machine learning gives one of the primary highlights for extricating knowledgefrom huge databasesfromendeavoursoperationaldatabases. Machine Learning in Medical Health Care is a developing field of exceptionally high significance for giving visualization and a more profound comprehension of medical data. Most machine learning methods depend on a set of features that define the behaviour of the learningalgorithm and directly or indirectlyinfluence the performance as well as the complexity of resulting models.

Heart disease and diabetes are two of the main sources of death everywhere throughout the world for a few years. There have been a few machine learning methods utilized for the conclusion of heart disease and diabetes previously. Neural Network, Logistic Regression Naïve Bayes etcetera are a portion of a couple of machine learning strategies utilized in the analysis of these diseases giving some measure of achievement. We explore various algorithms, for example, Neural Networks, K - Nearest Neighbours, Naive Bayes, Decision tree algorithms, Support vector classifiers and Logistic Regression alongside cross breed procedures including the above-utilized algorithms for the finding of heart disease and diabetes. The framework has been implemented in Python platform and prepared to utilize benchmark dataset from the UCI machine learning repository. The framework is likewise perhaps expandable for the new datasets.

Keywords: Naive Bayes, Decision Tree, Machine Learning, Logistic regression, KNN, SVM, Neural Network, Neurons

1 INTRODUCTION

The principle inspiration for doing this project is to introduce a prediction model for the prediction of the occurrence of diabetes and heart disease. Further, this un- dertaking work is pointed towards distinguishing the best classification method for recognizing the chance of heart disease or diabetes in a patient. This work is justified by playing out a similar report and analysis utilizing some machine learning algorithms for classification namely Naive Bayes, Decision Tree, K-Nearest Neighbors, Logistic Regression, Support Vector Classifier and Neural Networks. Despite the fact that these are normally utilized machine learning algorithms, disease prediction is a vital task including the highest possible exactness. Subsequently, the three algorithmic methods are assessed at various levels and sorts of classification strategies namely Naive Bayes, Logistic Regression and K-Nearest Neighbours has been studied previously but the accuracy of our Neural Network has been significantly better than these prior works. This will give scientists and clinical experts to set up a superior understanding and as- sist them with recognizing an answer for distinguish the best technique for anticipating heart illnesses as well as the chance of diabetes

The principle subject is prediction utilizing machine learning methods. Machine learning is generally utilized these days in numerous business applications like web-based business and some more. Prediction is one zone where this machine learning is utilized, our subject is about expectation of heart disease by handling a patient's dataset and an information of patients to whom we have to predict the opportunity of event of a heart illness and Diabetes.

The healthcare industry gathers a tremendous amount of human health information which, unfortunately, arenot "mined" todiscover the hidden data for successful decision making. The revelation of hidden patternsandrelationships regularly goesunexploited. Thehealthcare industry is still 'data-rich' but 'information poor. Thereis an abundance of information accessible inside the medicinal services frameworks. However, there is anabsence of successful investiga-

tion apparatuses to find hidden relationships in the information. Today medical administrations have made some amazing progress to treat patients with different diseases. Amongthe deadliest oneis the heart disease issue whichcan't be seenwith an unaided eye and comes in a flash when its limitations are reached. Today diagnosing patients accurately and regulating compelling medications have become a significant test. This area gives the details of the previous works and researches performed.

The significant challenge that the Healthcare business faces now-a-days is predominance of facilities. Diagnosing the illness accurately and giving compelling treatment to patients will characterize the nature of service. Poor diagnosis causes unfortunate results that are not acknowledged.

The framework examines the data in the medical field to evaluate the danger of disease. It utilizes methods to clean and change the data. Second, by utilizing different machine learning algorithms, it investigates the new approaching data point and orders the point into one of the two groups to be specific whether the individual is experiencing disease or not experiencing the disease. Different investigation procedures have been utilized to clean and change the data to fit the data into the machine learning model successfully. Contrasted with a few runs of the mill forecast algorithms, the expected accuracy of our proposed algorithm framework is the most elevated. The detailed process of how the data is cleaned and fitted in our machine learning model is discussed in Section 3 (Proposed Methodologies).

The proposed solution is to develop the machine learning model using different machine learning algorithms stated above. By the help of which the user can easily find out the risk of heart disease and diabetes by simply inputting some parameters from their medical report. The result of the analysis will draw the complete picture of whether the particular patient is having the disease or not which helps the doctor to diagnose the disease in the initial stages rather than believing in their intuition.

The essential point of this undertaking is to break down the "Pima Indian Diabetes Dataset" and "Heart Disease Dataset" and utilize Logistic Regression, Support Vector Machine, Naïve Bayes, K-Nearest Neighbors and Multi-Layer Perceptron (Neural Network) for forecast and build up an expectation motor and a straightforward UI which is simple and basic for new clients or patients to utilize. As far as we could know in the territory of clinical data analytics, none of the current work centres around the equivalent

2 LITERATURE SURVEY

The healthcare industry gathers a tremendous amount of human health informationthat, unfortunately, are not "mined" to discover the hidden data for successful decisionmaking. The revelation of hidden pattern and relationship regularly goes unexploited.The healthcare industry is still 'data-rich' but 'information poor'. There is abundance of information accessibleinside themedicial services framework. However, there is anabsence of successful investigation apparatuses to find hidden relationships in the information.Today medical administrations have made some amazing progress to treat patients with different diseases. Among the most deadly one is the heart disease issue whichcan'tbeseenwith an unaided eye and comes in a flash when its limitations are reached. Today diagnosing patients accurately and regulating compelling medicationshave become a significant test. This area gives the details of the previous works and researchs performed.

In the consent to the above assessment, there are different data mining systems that were used to assemble heart illness and diabetes. In the year 2000, look into coordinated byShusakuTsumoto says that as we individuals can't mastermind data if it is massive in size we ought to use the data miningtechniques that are available for finding different models from the open huge database and can be utilized again for clinical research and perform distinctive strategy on it.

Y. Alp Aslandogan, et. al. (2004), chipped away at three unique classifiers called K-closest Neighbor (KNN), Decision Tree, Naïve Bayesian, and utilized Dempster's' rule for these three perspectives to show up as one concluding choice. This order dependent on the combined thought shows increased accuracy.

Carlos Ordonez (2004), Assessed the risky to perceive and estimate the standard of relationship for heart disease. Dataset including clinical history of the patients having heart disease with the parts of hazard factors was gotten to by him, estimations of limited supply route and heart perfusion. Every one of these limitations were declared to contract the digit of structures, these are as per the following:

1. The features should seem on a oneside of the rule.
2. The rule should distinguish various features into the different groups.
3. The check of highlights accessible from the standard is composed by clinical history of individuals having heart disease as it were. The event or the non-appearance of heart diseases was anticipated by the creator in four heart veins with the two clus-ters of rules

S. Vijayarani, et. al. in (2013), utilized experimental outcomes completed utilizing divergent grouping strategies for heart disease dataset. The diverse classification frameworks which were utilized and tried by him are Decision Tree, Random Forest and LMT tree algorithm. WEKA device was utilized for examination.

Writer	Year	Methods/Techniques	Count ofattributes
Carlos et. al.	2001	Association Rule Mining	25
Latha et. al.	2008	Genetic Algorithm	14
		CANFIS	
Shantakumar et. al.	2009	MAFIA	13
		Clustering	
		K-Means	
Dr. K. Usha Rani	2011	Classification	13
		Neural Network	
Majabbar, et. al.	2011	Clustering	14
		Association Rule Mining	
		Sequence Number	
Nan-Chen, et. al.	2012	(EVAR)	
		Machine Learning	
		Markov Blanket	
Oleg, et. al.	2012	Artificial Neural Network	
		Genetic Polymorphisms	
Shadab, et. al.	2012	Navie Bayes	15
NidhiBhatla, et. al.	2012	Fuzzy Logic	4
		Weka Tool	
		Decision Tree	
		Naive Bayes	
		Classification via. Clustering	
JesminNahar,et. al.	2013	AprioriPredictive AprioriTertius	14
Ms. Ishtake, et. al.	2013	Decision Tree	15
		Neural Networks	
		Navie Bayes	
Ashish Kumar Sen1, et. al.	2013	Neuro-fuzzy	4
		Backpropagation Algorithm	
KanteshKumar Oad et. al.	2014	Fuzzy Rule Based Support System	6

Table 1: The different method/techniques used in the related prior works.

Year	Author	Purpose	Techniques used	Accuracy
2013	AbhishekTanej a	Heart disease prediction system using data mining techniques and different supervised Machine learning algorithms	J48	95.56%
			SMO	92.42%
			Multilayer perception	94.85%
2015	Priti Chandra et al.	Computational Intelligence Technique for early	Naïve Bayes	86.29%
2015	Cemil et al.	Propose application of knowledge discovering process on prediction of stroke patients	ANN	81.82% for training dataset 85.9% for test data set
			SVM	80.38% for train data set 84.26% for test data set
2016	Muhammad Saqlain et al.	Identification of Heart Failure by Using Unstructured	Logistic Regression	80.00%
			Neural Network	84.80%
			SVM	83.80%
			Random Forest	86.60%
			Decision Tree	86.60%
			Naïve Bayes	87.70%
2016	Marjia et al.	Heart disease prediction using WEKA tool and 10-Fold cross-validation	KStar	75%
			J48	86%
			SMO	89%
			Bayes Net	87%
			Multilayer Perceptron	86%
			Naïve Bayes	
2016	Dr. S. Seema et al.	Predict chronic disease by mining the data containing in historical health records	Decision tree Support Vector Machine (SVM)	Highest accuracy in case of heart disease 95.556% is achieved by SVM.

Table 2: The table showing the purpose, Technique used and the Accuracy of some prior research work.

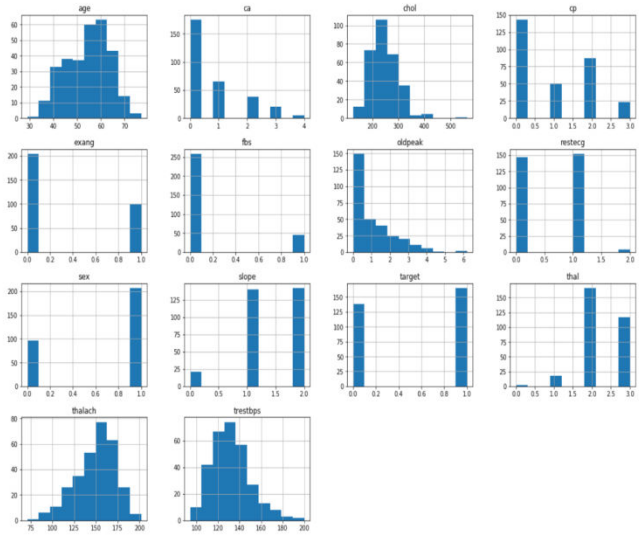
3 DATASETS AND PROPOSED METHODOLOGIES

3.1 Datasets

For this undertaking we have utilized The Cleveland Heart Dataset from the UCI Machine Learning Repository and the Pima Indians Diabetes Dataset as they are broadly utilized by the examples, plans and network.

A. The Cleveland Heart Dataset

The Cleveland heart dataset comprises 303 individual clinical reports in which 164 don't have any illness. In this dataset there are an aggregate of 97 female patients in which 25 individuals are the confirmed case, likewise there are 206 male patients in which 114 are determined to have the sickness. There are 6 missing values in this dataset and every single numeric value is perceived as numeric. We have 13 feature that are applicable to the particular infection with respect to the dataset are Age, Sex, Chest Pain Type, Resting Blood Pressure, Serum Cholesterol in mg/dl, Fasting Blood Sugar, Resting electrocardiographic result, Maximum heart rate achieved, Exercised - induced angina, Old peak, ST depression induced by exercise relative to rest, Number of major vessels coloured by fluoroscopy, Thal:3= Normal, 6=fixed defect, 7= reversible defect. The involvement



of each attribute for The Cleveland Heart Dataset is been studied and been shown as histogram in the below figure Fig. 1.

Figure 1: The involvement of each attribute with respect to number of instances for The Cleveland Heart Dataset as Histogram.

B. Pima Indians Diabetes Dataset

This dataset is initially from the National Institute of Diabetes and Digestive and Kidney Diseases.
The goal of the dataset is to analytically foresee whether a patient has diabetes, in view of certain diag-

nostic estimations included in the dataset. The dataset comprises 768 individual clinical reports in which 500 don't have any sickness. In this dataset all the patients are females of at least 21 years old of Pima Indian Heritage. The dataset consists of 8 features that are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Function, Age. The involvement of each attribute for Pima Diabetes dataset is been studied and been shown as histogram in the below figure Fig. 2.

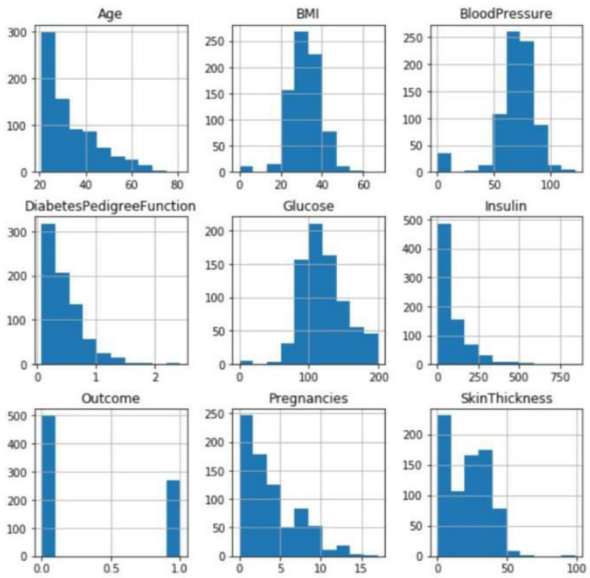


Figure 2: The involvement of each attribute with respect to number of instances for Pima Indians Diabetes Dataset as Histogram.

3.2 Proposed Methodologies

The block drawing for organization of heart disease and diabetes databank is shown in below figure Fig. 3.

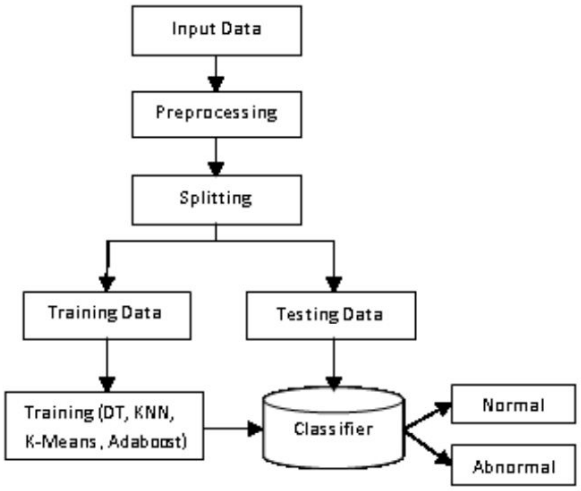


Figure 3: The organization of heart disease and diabetes databank.

A. Data Pre-processing

Cleaning: Data that we need to process won't be clean, that is it might contain noise or it might contain values missing from our process. We can't get great outcomes so to acquire great and immaculate outcomes we have to eliminate this, the process to take out this is data cleaning. We will fill missing qualities and can expel clamor by utilizing a few procedures like loading up with most normal qualities in missing spots. In this proposed framework the databasecontains NaN values. The NaN values can't be acecessed by the programming consequently these qualities need to change over into numerical qualities. In this methodology the mean of the section is determined and NaN values are replaced by the mean.

Transformation: This involves changing dataformat from one form to another that is making them most understandable by doing normalization, smoothing, and generalization, aggregation techniques on data.

Integration: Data that we need not process may not be from a single source once in a while it tends to be from various sources we don't incorporate them it might be an issue while expert-cessing so integration is one of the significant stages in information pre-processing and various issues are considered here to coordinate.

Reduction: At the point when we deal with information it might be difficult to understand and it might be hard to see once in a while so to make it understandable to the framework, we will diminish them to required design so we can accomplish great outcomes.

Splitting: The entire database is part into training and testing databases. The 80% data is taken for training while remaining 20 % data is utilized for testing.

B. Classification

The training data is trained by using six different machine learning algorithms i.e. Decision Tree, KNN, Naive Bayes, SVM, Neural Network and Logistic Regression. Each algorithm is explained in detail.

I. Naïve-Bayes Classification:

Naive Bayes classifiers are a gathering of straightforward probabilistic classifiers based by using Bayes theorem with strong (naive) freedom suppositions between the features. Naive Bayes classifiers are incredibly flexible by requiring different parameters directfor the number of features or pointers as a variable in a learning issue. It is the least perplexing and the snappi-

est probabilistic classifier, especially for the planning stage.

Naive Bayes classifier depends on Bayes theorem. This classifier utilizes restrictive autonomy wherein characteristic worth is autonomous of the estimations of different qualities. The Bayes theorem is as per the following:

Let X= x1, x2, ..., Xn be a lot of n qualities. In Bayesian learning, X is considered as proof and H be some speculation implies, the data of X has a place with explicit class C. We need to decide P (H|X), the likelihood that the speculation H holdsgiven proof, for example, data test X. As indicated by Bayes theorem the P (H|X) is communicated as:

$$P(H|X) = \frac{P(X|H) * P(H)}{P(X)}$$

- P(H|X) is the posterior probability of class (target) given predictor (attribute)
- P(H) is the prior probability of class.
- P(X|H) is the likelihood which is the probability of a predictor given class.
- P(X) is the prior probability of predictor

The flow chart for the implementation using Naïve-Bayes algorithm is as shown below in figure Fig. 4.

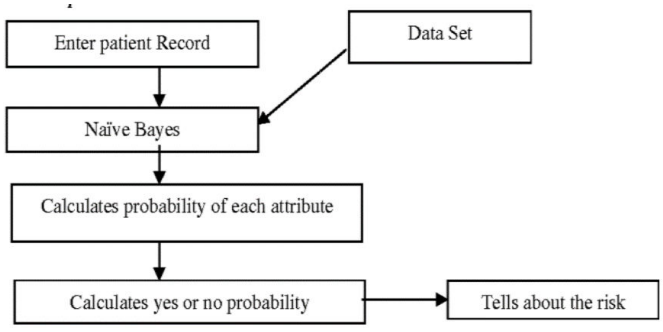


Figure 4: Flowchart of Naïve-Bayes algorithm.

II. Decision Tree:

Decision tree learning uses a decision tree as a prescient model which maps discernments about a thing to decisions about the thing's objective. It is one of the prescient demonstrating approaches used in estimations, data mining and Artificial Intelligence. Tree models where the target variable can take a limited arrangement of values arecalled characterization trees. In these tree structures, leaves address class stamps and branches address conjunctions of features that lead to those class names. Decision trees where the target variable can take ceaseless values (customarily real numbers) are called regression trees. In decision tree analysis, a decision tree can be used to apparently and explicitly address decisions and decision making. In data mining, as decision tree portrays data yet not decisions; rather the resulting characterization tree can be a commitment for decision making.

There are different sorts of decision trees. The only distinction is in logical standards that they use to top-notch the class of highlights through principle mining. An addition proportion decision tree is an extremely normal and productive category. It is the relationship among information gain and classified information.

In entropy framework, the trademark that decreases entropy and endeavors data gain is named as tree root. For choosing tree roots, it is first basic to appraise data increase everything being equal. Afterward, the property that misuses data increase will be selected.

$$\boldsymbol{E = -\sum_{i=1}^{k} P_i log_2 P_i}$$

Here k is count of response variable modules, piis-theratio offthe number of the ith class proceduresto a tot alcount of models

The analyzed Decision Tree for the Heart Prediction is shown in below figure Fig. 5.

This classifier settles on a decision tree reliant on which it gives out the class esteems to each datum point. Here, we can change the most extraordinary number of features to be thought of while making the model.

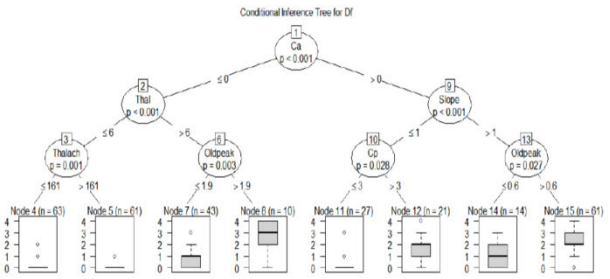


Figure 5: Decision Tree for Heart Prediction.

The flow chart for the implementation using Decision Tree algorithm is as shown below in figure Fig. 6.

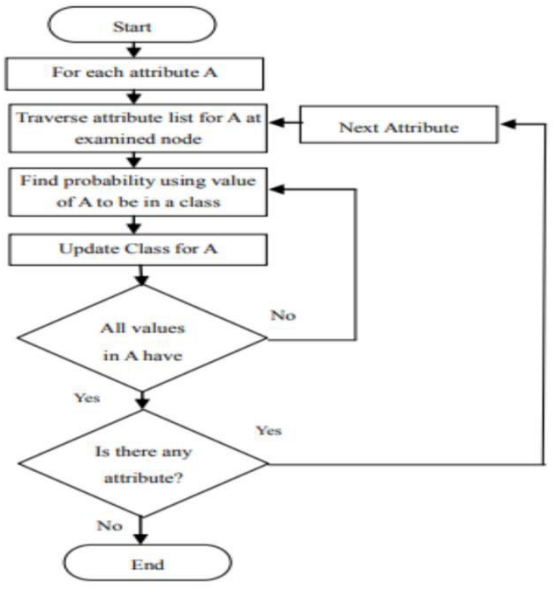


Figure 6: Flowchart of Decision Tree algorithm.

III. K-Nearest Neighbor (KNN):

K-Nearest Neighbor (KNN) is a straightforward, lazy and nonparametric classifier. KNN is favoured when all the highlights are consistent. KNN is additionally called case-based thinking and has been utilized in numerous applications like example acknowledgement, statistical estimation. Classification is acquired by distinguishing the nearest neighbour to decide the class of an obscure example. KNN is favoured over other classification algorithms because of its high combination speed and simplicity.

KNN classification has two steps:

1. Find the k number of instances in the dataset-that is closest to instanc e S
2. These k number of instances then vote to determine the class of ins tance S

The Accuracy of KNN relies upon separation metric and K esteem. Different methods of estimating the separation between the two cases are cosine, Euclidean separation. To assess the new obscure example, KNN processes its K nearest neighbours and doleout a class by greater part voting.

With the KNN algorithm, we have the opportunity to change the parameter'sweight. It implies that we may accept that a few parameters are more significant orhaving more effect than others. Among 8 parameters we use, we can classify themour data into 2 classifications, one is "non-clinical" parameters (Age and Sex) and theother is "clinical" parameters (CP, Trestbps, Trestbpd and so forth). We may feelthat clinical parameters are a higher priority than non-clinical, which we will see intest results. Alongside weighting, we should discover the estimation of "k" so it givesthe best classification result. Since it is a 2-decision classification ("yes") and ("No")k will be an odd number.

The flow chart for the implementation using KNN algorithm is as shown below figure Fig. 7.

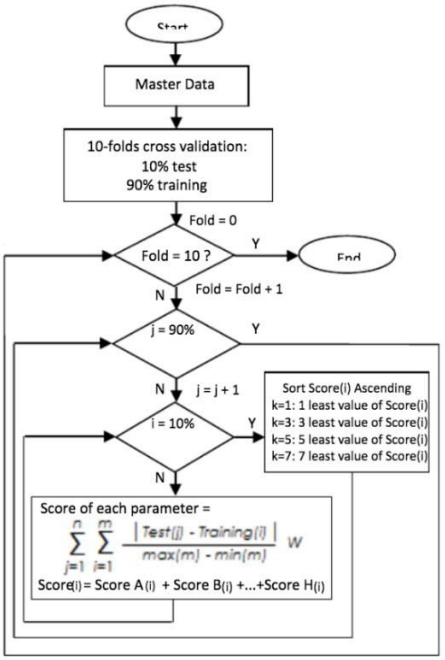


Figure 7: Flowchart of K-Nearest Neighbor algorithm.

IV. Logistic Regression:

Logistic Regression is a statistical analysis procedure that is utilized for foreseeing the data esteem dependent on the earlier perception of the data set. The logistic regression model predicts the needy data variable by examining the connection between at least one existing free factor. Logistic Regression is one of the significant instruments for forecast, which can likewise be utilized for characterizing and foreseeing the data dependent on the historical data. The actualized model is a twofold Logistic model that has subordinate factors with just two potential results i.e., one is a positive worth and another is the negative worth which is having 0 or 1 as a class mark.

It for the most part comprises of two significant stages: regularized cost work and regularized angle plummet. Cost Function is utilized for ascertaining the greatest probability estimation. Angle plummet is an iterative procedure for getting coefficients from preparing data. The procedure is rehashed until we get the ideal parameters of train data. The model is prepared with the ideal coefficient. Whenever a test data has been passed to the model dependent on the parameters can recognize whether the individual is having coronary illness or not, it tests the data utilizing the sigmoid capacity. The cost work is the technique that is utilized

for decreasing the mistakes of the anticipated name and the genuine mark. Slope plunge work is the technique that is utilized for computing the coefficient until we get a base estimation of the class mark.

- The cost function calculation is based on the formula given below:

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

Here,
m = number of instances; n = number of attributes; y = class label; x = train data features; θ = coefficient; λ =learning rate

- The gradient descent function:

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

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

- Generally,sigmoid function is used to map predictions to probability it is defined as:

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

Here,
x = test data features
θ = coefficients Whenever a test data is passed it calculates the value based on the parametersstored in the model. It calculates the probability of each class label. We return the maximum probability value of the class label xi.

The flow chart for the implementation using Logistic Regression algorithm is as shown below figure Fig. 8.

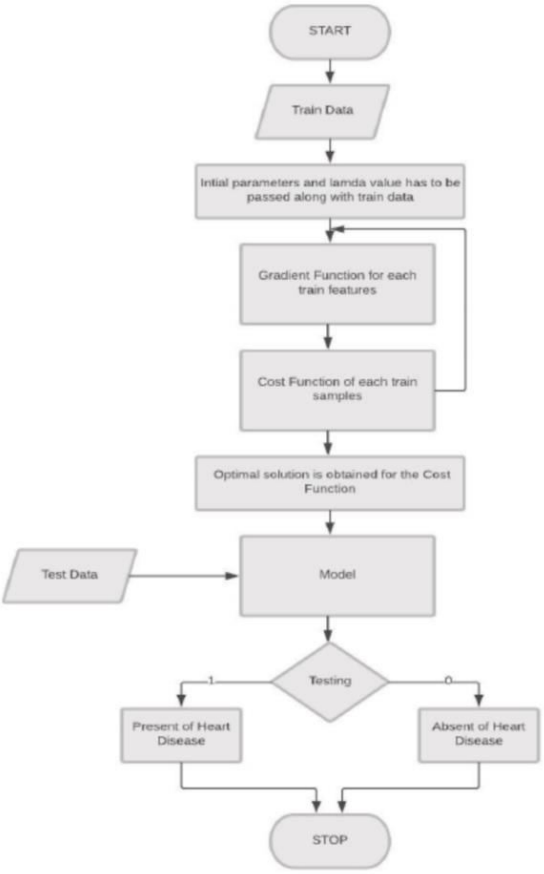


Figure 8: Flowchart of Logistic Regressionalgorithm.

V. Support Vector Machine (SVM):

It is based on the concept of decision planes that define decision boundaries. A decision plane is a hyperplane that separates the object shaving different class memberships. SVM classifiers separate the observations into two or more classes in such a way that maximum separation is achieved. A hypothetical hyperplane is the separator in SVM classification problems. In other words, SVM constructs a hyperplane that separates the two sets so as to minimize the number of misclassified points. Generally, there are two types of SVM models: linear and nonlinear. Linear SVM works better on linearly separable datasets but nonlinear SVM models work well even on hardly separable datasets. Since we are dealing with hardly separable data in our experiments, we use nonlinear SVM. The dual formulation of the nonlinear SVM function can be formulated as

$$MaxW(\alpha) = \sum_{i=1}^{m} \alpha_i - 0.5 \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j K(x_i.x_j)$$

Subject to,

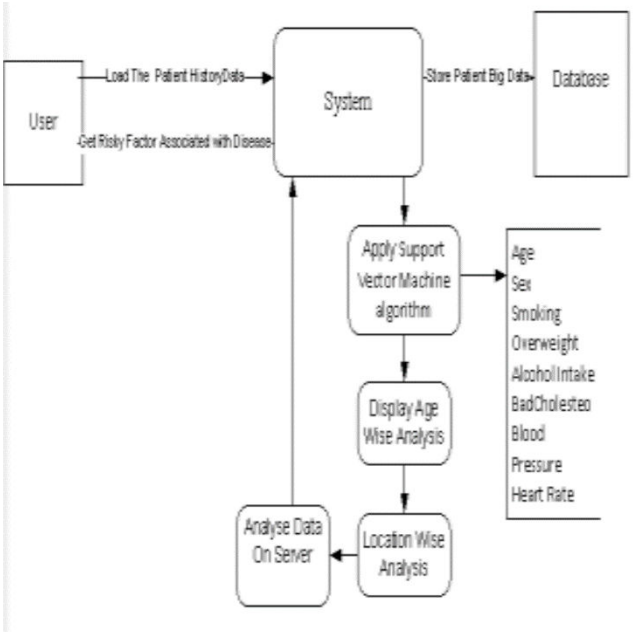
$$\sum_{i=1}^{m} \alpha_i y_i = 0,$$

Input vectors xi ∈Rm , i= 1, 2, 3, ..., m, which are called features or attributes are extracted from the database. Associated with every particular input we have a corresponding label (yi = ±1) which is called the target value or output in the database. The variable αi is the Lagrange multiplier in the dual formulation and C is a user-specified parameter representing the penalty for misclassification K (xi ,xj) is the kernel function and maps the original data points to another space. One of the popular choices for the kernel is Gaussian kernel which is also known as Radial Basis Function (RBF) in the literature. The formulation for this kernel is

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma}}$$

where parameter σ is known as the kernel width.

The flow chart for the implementation using SVM is as shown below figure Fig. 9.



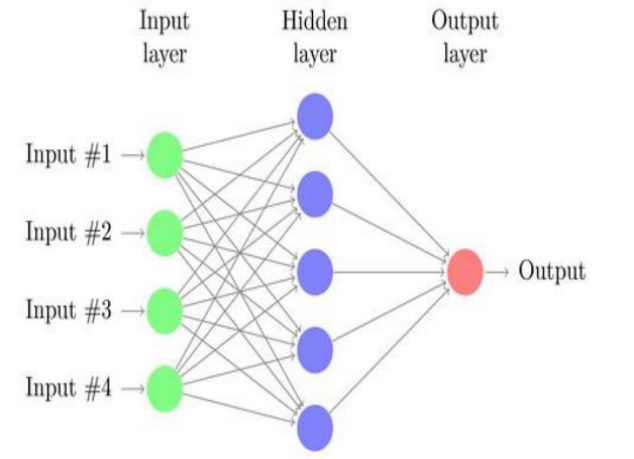


Figure 9: Flowchart of Support vector machinealgorithm.

VI. Neural Network:

The neural network is a computational model dependent on biological neural systems. Artificial neural nework (ANN) depends on perception of a human cerebrum. Human mind is an exceptionally confused web of neurons. Analogically ANN is an interconnected arrangement of three units, for example, input, hidden layer and yield units. In clinical determination, the patient's hazard factors or qualities are utilized as an input. The adequacy of artificial neural network was demonstrated in medication. ANN are utilized in predicting heart illness. Here the input layer comprising of 8 neurons compares to 8 significant characteristics. There is one output class variable which takes the worth either 0 or 1. The worth 0 speaks to that the individual isn't experiencing heart disease and the worth 1 speaks to that the individual experiences heart disease. The quantity of hubs utilized in the hidden layer are 3. The Sample Artificial Neural Network is shown in Fig. 10 below.

Figure 10: Simple Artificial Neural Network.

The primary advantages of neural network systems are high exactness. The uses of neural systems are accounting tallying, medication, misrepresentation identification and so on. In light of the learned system or training dataset, the neural system can foresee the nearness or nonappearance of heart disease for the testing dataset.

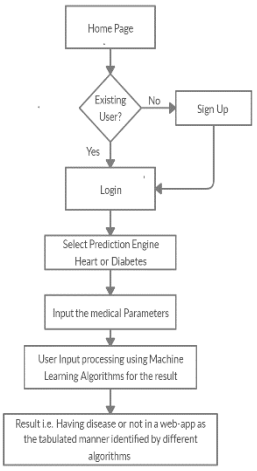
The perfomance proportions of neural system are calculated utilizing different estimates, for example, accuracy, specificity and sensitivity.

Accuracy = (TP+TN) / (TP+TN+FP+FN)
Specificity = TN / (TN+FP)
Sensitivity = TP / (TP+FN)

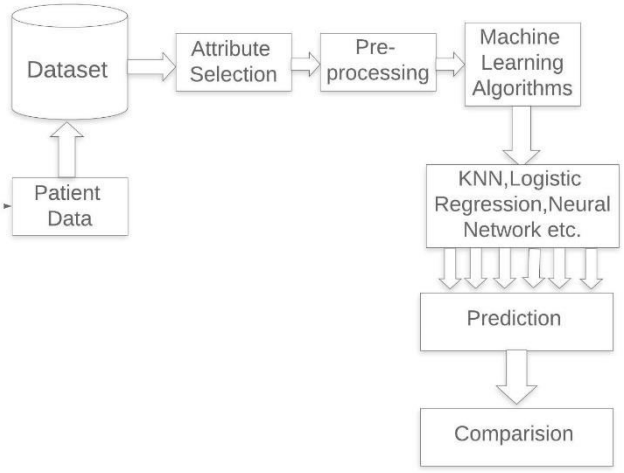
Where ,

- TP = True Positive; that is the quantity of tests which are delegated having heart illness while they really have heart illness.
- TN = True Negative; that is the quantity of tests which are named not having heart illness while they were really not.
- FN = False Negative; that is the quantity of tests which are named not having heart illness while they were really have heart illness
- FP = False Positive; that is the quantity of tests which are delegated having haert illness while they were really not.

The flow chart for the implementation using Neural Network algorithm is as shown below figure Fig.



11.



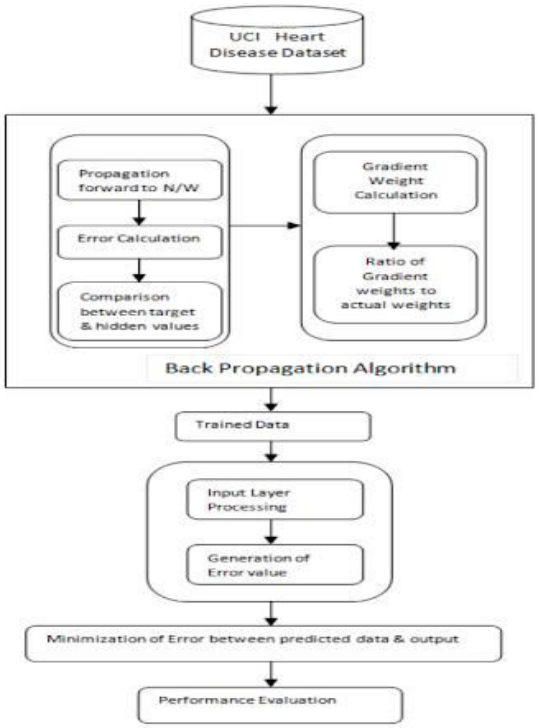


Figure 11: Flowchart of Artificial Neural Network.

4 PROPOSED SOLUTION

By the above experiment what we say is different machine learning algorithms may produce different results. So, for every prediction, we need to have a comparison of all the algorithms to get accurate results whereas by using just one algorithm we may not obtain the clear picture of the presence of disease. So, it's better to have a combination of algorithms like k-means, logistic regression, SVC, KNN, neural networks, decision tree and Naïve Bayes to get more a clearer pictureof the presence of disease. In the experiment carried out, even though the highest accuracy is obtained by multilayer perceptron neural network algorithm but we still try to show the results using the combination of all the algorithms mentioned above so as to give a better idea to the presence of disease to the end user.

Figure 12: The implementation diagram

The algorithms which are implemented for this project have been depicted in the below table 3 for both diabetes and Heart disease with their respective accuracies.

Machine Learning Algorithm	Heart Disease Accuracy	Diabetes Accuracy
Logistic Regression	81.57 %	76.01 %

SVC	85.52 %	80.20 %
K-Nearest Neighbour	71.73 %	66.37 %
Neural Network	97.00 %	91.70 %
Naïve Bayes	80.00 %	73.00 %
Decision Tree	80.43 %	74.13 %

Table 3: The accuries achieved by our model using different machine learning algorithms.

Figure 13; The flowchart for our UI system.

The system proposed uses the web-app as the frontend application for the user or client which is been build using Django framework having the feature of login and storing the data in the local database. The user first directed to the home page then the new user can create the account and the existing one can login into the system. After the successful login into the system the user is provided with the choice to choose the prediction engine i.e. Diabetes or Heart Disease which to be predicted . After the choice is made user will be directed to the form page which takes the value of medical report parameters as a user input then the provided user input is been processed and the result is shown indicating whether the user having the risk of disease or not indicated against the six machine learning algoritms used in this study. The implementation diagram is shown is figure Fig. 12 and the frontend flowchart is depicted in figure Fig. 13 above.

5 CONCLUSIONS

In the above paper we have contemplated different classification algorithms that can be utilized for grouping of heart illness and diabetes datasets. We have seen various methods that can be utilized for order and the exactness obtained by them. This examination informs us regarding divergent innovations that are utilized in disparate papers with a unique tally of qualities with various correctness'srelying upon the devices intended formexecution. The precision of the structure can be additionally redesigned by making different blends of information mining methods and by parameter tuning too.

The project undertook the study of various algorithms that include Neural Network, Naive Bayes, KNN, SVC, Decision tree and Logistic Regression that can be effectively implemented in Python to predict the heart attacks and diabetes. A couple of algorithm calculations, for example, neural systems perstructure a vastly improved investigation and give preferred performance over different examination papers which have just been refered to.

As is realized that heart illness has an intricate pathology, in this manner, the improvement of our model despite everything needs guidance and recommendations from a specialist and including a few credits of patient information to decide the in-strained quality of heart illness whether heart illness is as of now at the degree of incessant or not yet. The project can be utilized as a significant device for specialists and wellbeing specialists to predict certain critical cases in the training and used to prompt the patient as needs be. Any nonmedical employee can utilize this product and anticipate the heart disease and lessen the time unpredictability of the specialists.

6 REFERENCES

1. V. Krishnaiah, G. Narasimha, N. Subhash Chandra, "Heart Disease Prediction System using Data Mining Techniques and Intelligent Fuzzy Approach: A Review" IJCA 2016.
2. Rajesh, T Maneesha ,Shaik Hafeez, Hari Krishna "Prediction of Heart Disaease using Machine learning Algorithms".
3. S. Vijayarani, S Sudha "A study of Heart Disease Prediction in Data Mining"
4. Avinash Golande, Pavan Kumar T "Heart Disease Prediction Using Effective Machine Learning Techniques"
5. K. Uma Maheswari "Neural Network based Heart Disease Prediction" Anna University.
6. Adil Hussain She, Dr. Pawan Kumar Chaurasia "A rieview on heart disease prediction using machine learning techniques"
7. C. Beulah Christalin, LathaS. CarolinJeeva "Improving the accuracy of prediction of heart dis ease risk based on ensemble classification techniques"
8. A H Chen, 2011. HDPS: heart disease prediction system; 2011 computing in cardiology
9. Dinesh Kumar G, 2018.Prediction of cardiovascular disease using machine learning algorithms, proceeding 2018 IEEE International Conference on Current Trends toward Converging Technologies, Coimbatore, India.
10. Carlos Ordonez, Association Rule Discovery with the Train and Test Approach for Heart Disease Prediction", IEEE Transactions on Information Technology in Biomedicine, Vol. 10, No. 2, April 2006.
11. Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors ", Proceedings of IEEE Conference on Information & Communication Technologies, 2013.
12. Sikander Singh Khurl, Gurpreet Singh, Ranking Early Signs of Coronary Heart Disease

Among Indian Patients", IEEE International Conference on Computing for Sustainable Global Development, 2015

13. DeeannaKelley "Heart Disease: Causes, Prevention, and Current Research" in JCCC Honors Journal
14. Nabil Alshurafa, Costas Sideris, Mohammad Pourhomayoun, Haik Kalantarian, Majid Sarrafzadeh "Remote Health Monitoring Outcome Success Prediction using Baseline and First Month Intervention Data" in IEEE Journal of Biomedical and Health Informatics
15. Ponrathi Athilingam, Bradlee Jenkins, Marcia Johansson, Miguel Labrador "A MobileHealth Intervention to Improve Self-Care inPatients with Heart Failure: Pilot Randomized Control Trial" in JMIR Cardio 2017, vol. 1, issue 2.

.
.