Disease Prediction Using Machine Learning

[1]Khushal Kailash Hinduja, [2]Aejazul Khan

Thadomal Shahni College of Engineering
Mumbai – 400050 (Maharashtra), India

Abstract- **Machine Learning is used across many ranges around the world. The healthcare industry is no exclusion. Machine Learning can play an essential role in predicting presence/absence of locomotors disorders, Heart diseases and more. Such information, if predicted well in advance, can provide important intuitions to doctors who can then adapt their diagnosis and dealing per patient basis. We work on predicting possible Heart Diseases and Diabetic Risk in people using Machine Learning algorithms. In this project we perform the comparative analysis of classifiers like Decision Tree, Logistic Regression, Random Forest and Support Vector Machine and we propose an ensemble classifier which perform hybrid classification by taking strong and weak classifiers since it can have multiple number of samples for training and validating the data.**

I. INTRODUCTION

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. Heart disease is one of the biggest causes of morbidity and mortality among the population of the world. Prediction of cardiovascular disease is regarded as one of the most important subjects in the section of data analysis. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduces the complications. Machine learning proves to be effective in assisting in making decisions and predictions from the large quantity of data produced by the health care industry. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithm. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can say that this technique can be very well adapted to do the prediction of heart disease.

II. OBJECTIVES

The primary goal is to develop a prediction engine which will allow the users to check whether they have heart disease sitting at home. The user need not visit the doctor unless he has heart disease, for further treatment. Doctors rely on common knowledge for treatment. When common knowledge is lacking, studies are summarized after some number of cases have been studied. But this process takes time, whereas if machine learning is used, the patterns can be identified earlier. For using machine learning, a huge amount of data is required. There is very limited amount of data available depending on the disease. Also, the number of samples having no diseases is very high compared to the number of samples having the disease.

The primary aim of this project is to analyze the data and use Support Vector Machine for prediction. The secondary aim is to develop a web application that allows users to predict heart disease utilizing the prediction engine.

III. LITERATURE SURVEY

With growing development in the field of medical science alongside machine learning various experiments and researches has been carried out in these recent years releasing the relevant significant papers.

[1]One paper has training and testing of dataset performed by using neural network algorithm multi-layer perceptron. In this algorithm there will be one input layer and one output layer and one or more layers are hidden layers between these two input and output layers. Through hidden layers each input node is connected to output layer. This connection is assigned with some random weights. The other input is called bias which is assigned with weight based on requirement the connection between the nodes can be feedforwarded or feedback.

The following give an overview of the various methodologies used by various authors for disease prediction using machine learning methodologies. We can observe that there is fine comparison made between machine learning algorithms whether they are able to predict the presence of the disease with a greater accuracy, achieving optimal performance. The research efforts presented by the authors in the following papers are focused in developing and evaluating a web-based tool for disease prediction. Heart disease is even being highlighted as a silent killer which leads to the death of a person without obvious symptoms. The nature of the disease is the cause of growing anxiety about the disease & its consequences. Hence continued efforts are being done to predict the possibility of this deadly disease in prior. So that various tools & techniques are regularly being experimented with to suit the present-day health needs. Machine Learning techniques can be a boon in this regard. Even though heart disease can occur in different forms, there is a common set of core risk factors that influence whether someone will ultimately be at risk for heart disease or not. By collecting the data from various sources, classifying them under suitable headings & finally analysing to extract the desired data we can conclude. This technique can be very well adapted to the do the prediction of heart disease. As the well-known quote says "Prevention is better than cure", early prediction & its control can be helpful to prevent & decrease the death rates due to heart disease.

IV. DOMAIN EXPLANATION

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

Types of machine learning:

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches:

Supervised learning: In this type of machine learning, data scientists supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.

Unsupervised learning: This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined. 4

Semi-supervised learning: This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled training data, but the model is free to explore the data on its own and develop its own understanding of the data set.

Reinforcement learning: Data scientists typically use reinforcement learning to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way.

The type of algorithm data scientists choose to use depends on what type of data they want to predict.

The classification can be done by various methods as listed below:

1. SUPPORT VECTOR MACHINE

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine. Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. The followings are important concepts in SVM –

Support Vectors - Data Points that are closest to the hyperplane are called support vectors. Separating line will be defined with the help of these data points.

2. DECISION TREE

Decision Tree is a Supervised learning technique that can be used for both classification and regression problems, but mostly it is preferred for solving classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision Tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a Decision Tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm. A Decision Tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees. The Decision Tree Algorithm belongs to the family of supervised machine learning algorithms. It can be used for both a classification problem as well as for a regression problem. The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree.

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision Tree:

- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure. In Decision Tree the major challenge is to identify the attribute for the root node in each level. This process is known as attribute selection.

We have two popular attribute selection measures:

1. Information Gain: When we use a node in a Decision Tree to partition the training instances into smaller subsets, the entropy changes. Information gain is a measure of this change in entropy. Entropy is the measure of uncertainty of a random variable, it characterizes the impurity of an arbitrary collection of examples. The higher the entropy the more the information content.

2. Gini Index: Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified. It means an attribute with lower Gini index should be preferred. Sklearn supports “Gini” criteria for Gini Index and by default, it takes “gini” value.
3. Classification and Regression Tree (CART): It is a dynamic learning algorithm which can produce a regression tree as well as a classification tree depending upon the dependent variable. Working: In a Decision Tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of the root attribute with the record attribute and, based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and moves further. It continues the process until it reaches the leaf node of the tree.

3. RANDOM FOREST

Random Forest is a supervised learning algorithm. It is an extension of machine learning classifiers which include the bagging to improve the performance of Decision Tree. It combines tree predictors, and trees are dependent on a random vector which is independently sampled. The distribution of all trees are the same. Random Forests splits nodes using the best among of a predictor subset that are randomly chosen from the node itself, instead of splitting nodes based on the variables. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest consists of trees. It is said that the more trees it has, the more robust a forest is. Random Forests create Decision Trees on randomly selected data samples, get predictions from each tree and select the best solution by means of voting. It also provides a pretty good indicator of the feature importance. Random Forests have a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset. Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

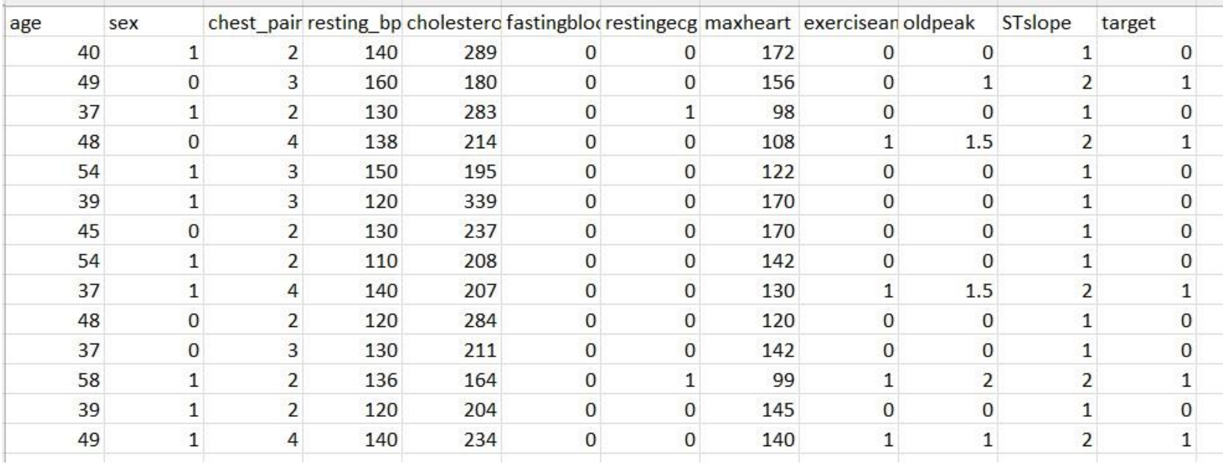
4. LOGISTIC REGRESSION

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1. Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas logistic regression is used for solving the classification problems. In Logistic regression, instead of fitting a regression line, we fit an "S"shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc. Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. This algorithm allows models to be updated easily to reflect new data, unlike Decision Tree or Support Vector Machine. The update can be done using stochastic gradient descent. Logistic Regression outputs well-calibrated probabilities along with classification results. This is an advantage over models that only give the final classification as results. If a training example has a 95% probability for a class, and another has a 55% probability for the same class, we get an inference about which training examples are more accurate for the formulated problem.
We will use the Support Vector Machine algorithm because various researches prove it is one of the best algorithms to give good performance.
Web development refers to the building, creating, and maintaining of websites. It includes aspects such as web design, web publishing, web programming, and database management. It is the creation of an application that works over the internet i.e. websites.
Web Development can be classified into two ways – Frontend and Backend. We have used HTML, CSS, Javascript and Bootstrap for frontend and PHP, Java, and python for backend.

V. RELATED WORK

A. Dataset

We have used two dataset of patients for predicting as we are predicting two different diseases – Heart and Diabetes.
Heart Disease Dataset consists of data of 1190 patients with 12 attributes. There are 11 input attributes and one attribute is used for predicting.



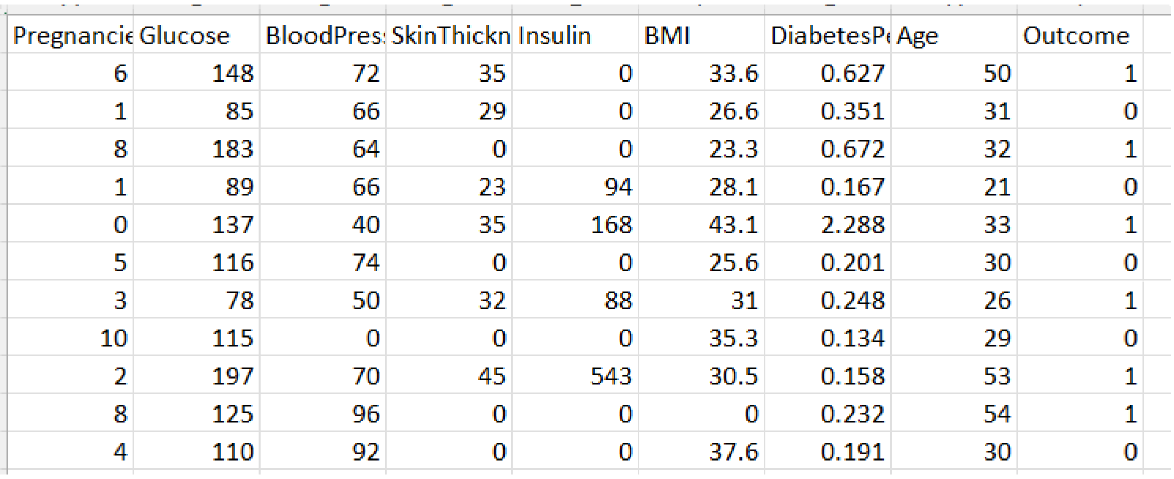
age	sex	chest_pair	resting_bp	cholesterc	fastingbloc	restingecg	maxheart	exercisean	oldpeak	STslope	target
40	1	2	140	289	0	0	172	0	0	1	0
49	0	3	160	180	0	0	156	0	1	2	1
37	1	2	130	283	0	1	98	0	0	1	0
48	0	4	138	214	0	0	108	1	1.5	2	1
54	1	3	150	195	0	0	122	0	0	1	0
39	1	3	120	339	0	0	170	0	0	1	0
45	0	2	130	237	0	0	170	0	0	1	0
54	1	2	110	208	0	0	142	0	0	1	0
37	1	4	140	207	0	0	130	1	1.5	2	1
48	0	2	120	284	0	0	120	0	0	1	0
37	0	3	130	211	0	0	142	0	0	1	0
58	1	2	136	164	0	1	99	1	2	2	1
39	1	2	120	204	0	0	145	0	0	1	0
49	1	4	140	234	0	0	140	1	1	2	1

Fig 1 Heart Disease Dataset

Sr. No.	Attribute	Description	Type
1	Age	Patient's age	Numerical
2	Sex	Gender of patient (male-0 female-1)	Nominal
3	Chest_pain	Chest pain type	Nominal
4	Resting_bp	Resting blood pressure(in mm Hg on admission to hospital)	Numerical
5	Chol	Serum cholesterol in mg/dl	Numerical
6	fastingblood	Fasting blood sugar>120 mg/dl, true-1 false-0)	Nominal
7	Restingecg	Resting electrocardiographics result (0 to 1)	Nominal
8	maxheartrate	Maximum heart rate Achieved	Numerical
9	Exang	Exercise included agina (1-yes 0-no)	Nominal
10	Oldpeak	ST depression introduced by exercise relative to rest (0 to .2)	Numerical
11	Slope	The slop of the peak exercise ST segment (0 to 1)	Nominal
12	Targets	1 or 0	Nominal

Table 1 Heart Disease Dataset Attributes

Diabetes Disease Dataset consists of data of 768 patients with 9 attributes. There are 8 input attributes and 1 attribute is used for predicting.



Pregnancie	Glucose	BloodPres	SkinThickn	Insulin	BMI	DiabetesPe	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0

Fig 2 Diabetes Disease Dataset

Sr. No.	Attribute	Description	Type
1	Pregnancies	If pregnant then the month	Numerical
2	Glucose	Sugar Levels current	Nominal
3	Bloodpressure	Blood Pressure levels on admission to hospital or checkup	Nominal
4	Skinthickness	Thickness of how fat the skin layer is of the patient	Numerical
5	Insulin	Dosage of insulin if any	Numerical
6	BMI	Body mass index is a value derived from the mass and height of a person.	Nominal
7	Diabetespedigreefunction	Assumption of diabetes as per previous history of family.	Nominal
8	Age	Patient’s age	Numerical
9	Outcome	1 or 0	Nominal

Table 2 Diabetes Disease Dataset Attributes

B. Data Preparation

- Data Pre-processing and Cleaning: Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. Data pre-processing is a proven method of resolving such issues. The purpose of pre-processing is to convert raw data into a form that fits machine learning. Structured and clean data allows a data scientist to get more precise results from an applied machine learning model. The technique includes data formatting, cleaning, and sampling. Data formatting. The importance of data formatting grows when data is acquired from various sources by different people. This set of procedures allows for removing noise and fixing inconsistencies in data. A data scientist can fill in missing data using imputation techniques, e.g. substituting missing values with mean attributes. A specialist also detects outliers — observations that deviate significantly from the rest of distribution. If an outlier indicates erroneous data, a data scientist deletes or corrects them if possible. This stage also includes removing incomplete and useless data objects.

- Data Splitting: A dataset used for machine learning should be partitioned into three subsets — training, test, and validation sets.

○ Training set: A data scientist uses a training set to train a model and define its optimal parameters — parameters it must learn from data. Test set.

○ A test set is needed for an evaluation of the trained model and its capability for generalization. The latter means a model’s ability to identify patterns in new unseen data after having been trained over a training data. It is crucial to use different subsets for training and testing to avoid model overfitting, which is the incapacity for generalization we mentioned above.

Validation set. The purpose of a validation set is to tweak a model's hyperparameters — higher-level structural settings that cannot be directly learned from data. These settings can express, for instance, how complex a model is and how fast it finds patterns in data. The proportion of a training and a test set is usually 80 to 20 percent, respectively.

○ A training set is then split again, and its 20 percent will be used to form a validation set. At the same time, machine learning practitioner Jason Brownlee suggests using 66 percent of data for training and 33 percent for testing. A size of each subset depends on the total dataset size.

The more training data a data scientist uses, the better the potential model will perform. Consequently, more results of model testing data lead to better model performance and generalization capability.

- Modelling After pre-processing the collected data and split it into three subsets, we proceed with a model training. This process entails "feeding" the algorithm with training data. An algorithm will process data and output a model that is able to find a target value in new data. Two model training styles are most common — supervised and unsupervised learning. The choice of each style depends on whether you must forecast specific attributes or group data objects by similarities.

- Model Deployment: Deployment is the method by which we integrate a machine learning model into an existing production environment to make practical business decisions based on data. It is one of the last stages in the machine learning life cycle and can be one of the most cumbersome. Often, an organization's IT systems are incompatible with traditional model-building languages, forcing data scientists and programmers to spend valuable time and brainpower rewriting them.

The working of the system starts with the collection of data and selecting the important attributes. Then the required data is pre-processed into the required format. The data is then divided into two parts training and testing data. The algorithms are applied and the model is trained using the training data. This system is implemented using the following modules.

1.) Collection of Dataset
2.) Selection of attributes
3.) Data Pre-Processing
4.) Balancing of Data
5.) Disease Prediction

C. Architecture Diagram

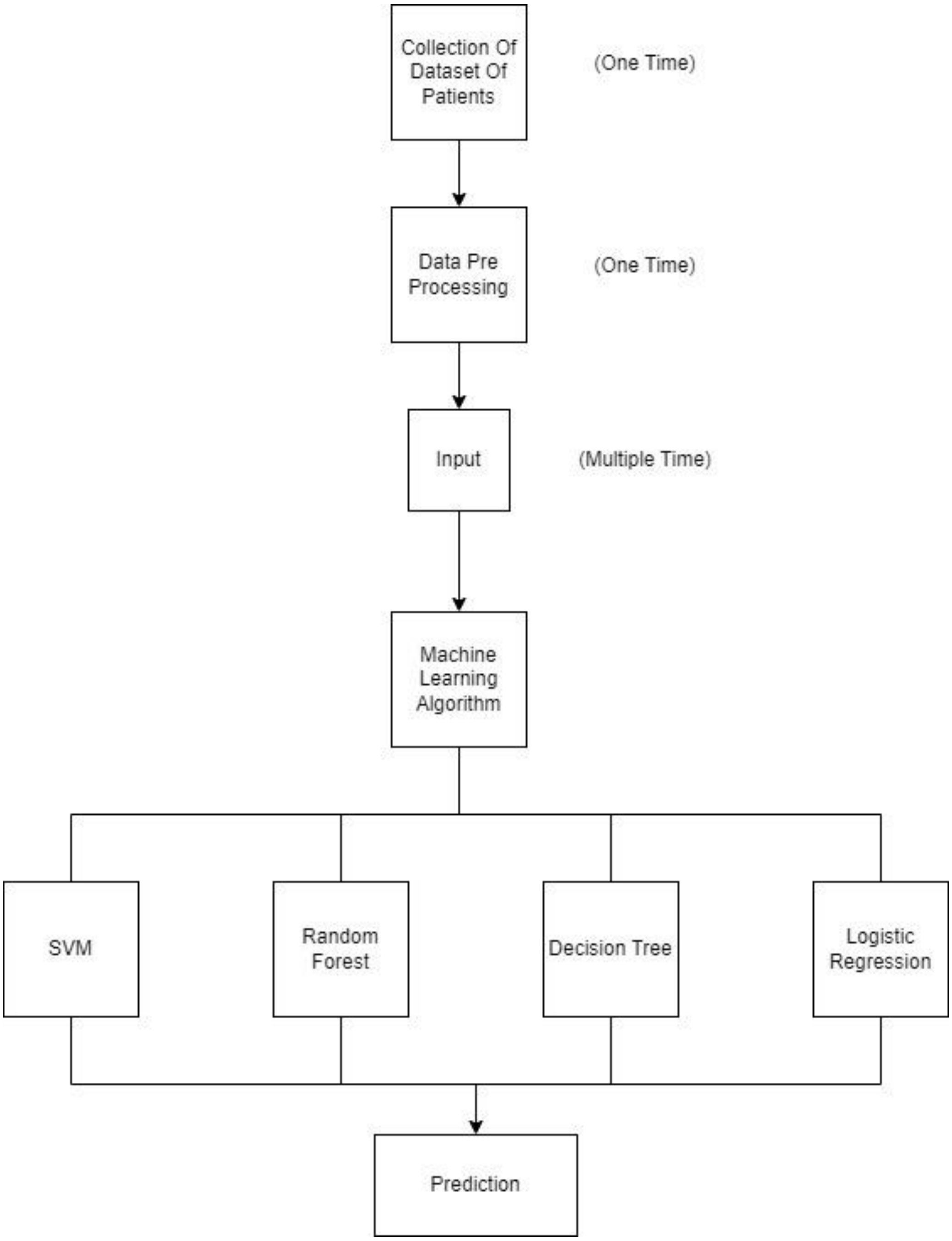


Figure 3 Architecture Diagram

As shown in above architecture diagram, we need to first collect the data for patients who are facing problems like heart disease and diabetes which is a one time process. We need to even consider those well and fit as this helps in building accuracy on the project. From the dataset collected, we need to filter out the values that are missing or any value that is not correctly entered. This step is called as data – preprocessing and is necessary for accuracy and is performed once. From the processed data, we need to build a Machine Learning model which will help us to make predictions. There are various Machine Learning algorithms which can be used, we have used four algorithms in the project which are – Support Vector Machine, Random Forest, Decision Tree, and Logistic Regression. Using these models, we have trained the model and build a model for making predictions on the disease. The user can input the details in the form provided to get predictions as per different models of machine learning. Models are created once and same are used for predicting with multiple conditions.

E. Performance Metrics

In this project, various machine learning algorithms like SVM, DecisionTree, Random Forest, Logistic Regression are used to predict heart disease and diabetes disease. Various attributes of the patient like gender, chest pain type, fasting blood pressure, serum cholesterol, exang, etc are considered for this project. The accuracy for individual algorithms has been measured with respect to data and whichever algorithm is giving the best accuracy, that is considered for the heart disease prediction. For evaluating the experiment, various evaluation metrics likeaccuracy, confusion matrix, precision, recall, and f1-score are

considered.
Accuracy- Accuracy is the ratio of the number of correct predictions to the totalnumber of inputs in the dataset. It is expressed as:
Accuracy = (TP + TN) /(TP+FP+FN+TN)

Confusion Matrix- It gives us a matrix as output and gives the total performance of the system.

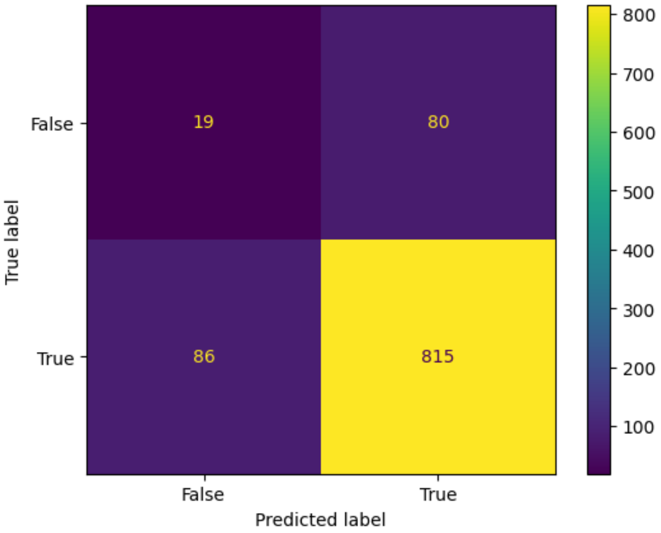


Fig 3 Heart Disease Dataset Confusion Matrix

From the above confusion matrix, we can conclude that the dataset which we have chosen for heart disease prediction, only 815 patients are predicted correctly by the machine. 80 patients are predicted incorrectly by the model, 19 are correctly predicted, and 86 are predicted incorrectly.

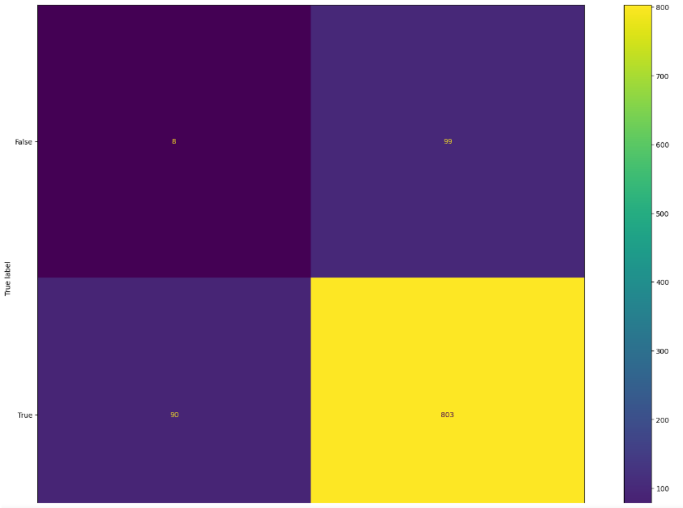


Fig 4 Diabetes Disease Dataset Confusion Matrix

From the above confusion matrix, we can conclude that the dataset which we have chosen for diabetes disease prediction, only 803 patients are predicted correctly by the machine. 99 patients are predicted incorrectly by the model, 08 are correctly predicted, and 90 are predicted incorrectly.
Where,
TP: True positive
FP: False Positive
FN: False Negative
TN: True Negative

Correlation Matrix: The correlation matrix in machine learning is used for featureselection. It represents dependency between various attributes.

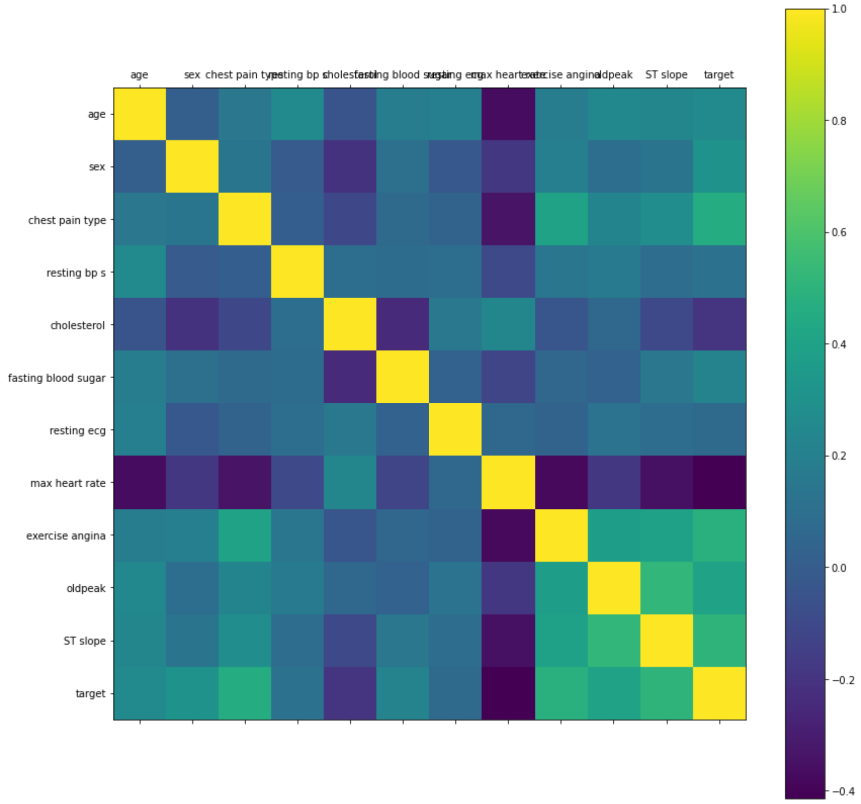


Fig 5 Heart Disease Dataset Correlation Matrix

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.

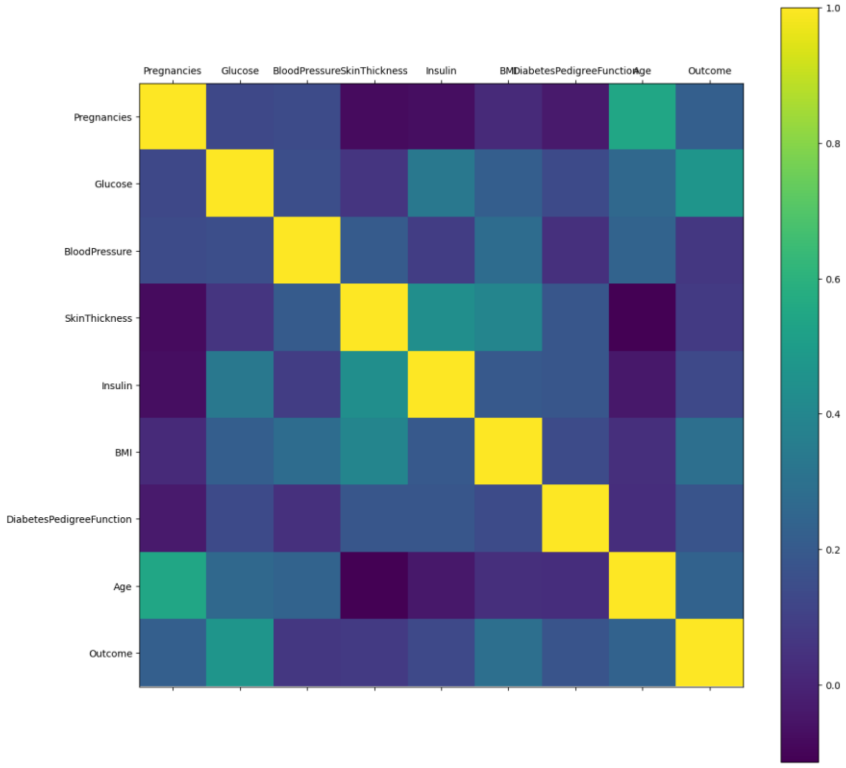


Fig 6 Diabetes Disease Dataset Correlation Matrix

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table. It is a powerful tool to summarize a large dataset and to identify and visualize patterns in the given data.

VI. RESULTS

After performing the machine learning approach for training and testing we find that accuracy of the Random Forest is better compared to other algorithms for Heart Disease and accuracy of Logistic Regression is better as compared to other algorithms for Diabetes Disease. Accuracy is calculated with the support of the confusion matrix of each algorithm, here the numbercount of TP, TN, FP, FN is given and using the equation of accuracy.

Algorithm	Heart Disease	Diabetes Disease
SVM	81.5%	75.3%
Logistic Regression	81.09%	76.6%
Decision Tree	84.1%	72.6%
Random Forest	91.6%	74.6%

Table 3 Accuracy

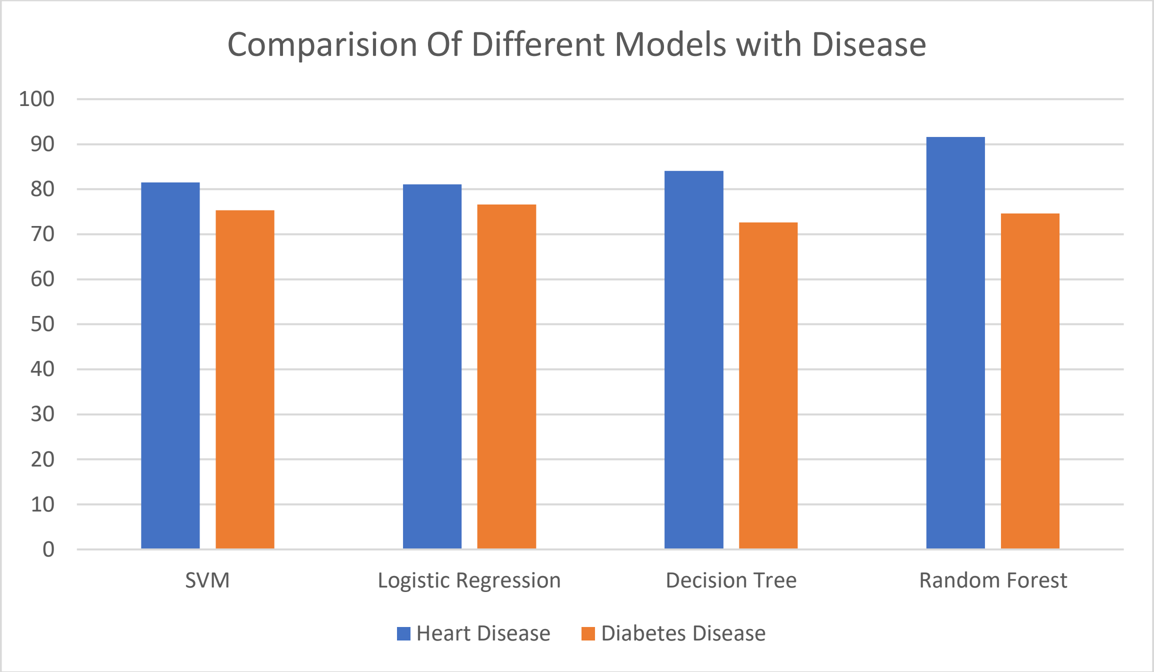


Figure 7 Comparision Chart

In the above graph, after implementing and testing the models, we found that the accuracy of the Random Forest for Heart Disease Prediction is good and Logistic Regression for Diabetes Prediction. These models can be used for better accuracy and results for the patients. We have compared the results for different algorithms with respect to heart disease and diabetes prediction.

VII. Conclusion

Heart diseases are a major killer in India and throughout the world, application of promising technology like machine learning to the initial prediction of heart diseases will have a profound impact on society. The early prognosis of heart disease can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. The number of people facing heart diseases is on a raise each year. This prompts for its early diagnosis and treatment. The utilization of suitable technology support in this regard can prove to be highly beneficial to the medical fraternity and patients. In this paper, the four different machine learning algorithms used to measure the performance are SVM, Decision Tree, Random Forest, Logistic Regression applied on the dataset.

All the four machine learning methods accuracies are compared based on which one prediction model is generated. Hence, the aim is to use various evaluationmetrics like confusion matrix, accuracy, precision, recall, and f1-score which predicts the disease efficiently.

VIII. Future Work

At some point in future, the machine learning model will make use of a larger training dataset, possibly more than a million different data points maintained in electronic health record system. Although it would be a huge leap in terms of computational power and software sophistication but a system that will work on artificial intelligence might allow the medical practitioner to decide the best suited treatment for the concerned patient as soon as possible. A software API can be developed to enable health websites and apps to provide access to the patients free of cost. The probability prediction would be performed with zero or virtually no delay in processing.

Data for more patients needs to be collected as that is the main reason the accuracy is not up to the mark. Since this is related to health so testing should be intensive and should be consulted with doctors.

We can add more diseases to be predicted using this website. This will help the user to get more accurate reports in one platform and also consultation can be added.
A facility to book appointments should also be there and features such as video consultation can also be added after predicting the results which should work 24 X 7 as health emergency can come at any given time.

REFERENCES:

1. Aditi Gavhane , "Heart Disease Prediction using Machine Learning ", IJERT, Conference Paper , 2021
2. Bilal A. Mateen, James Liley, Alastair K. Denniston, Chris C. Holmes and Sebastian J. Vollmer, "Improving the quality of machine learning in health applications and clinical research"", IJERT, Conference Paper , 2020
3. Amir Masoud Rahmani , Efat Yousefpoor , Mohammad Sadegh Yousefpoor , Zahid Mehmood, Amir Haider , Mehdi Hosseinzadeh , and Rizwan Ali Naqvi "Machine Learning (ML) in Medicine: Review, Applications, and Challenges", MDPI, Review Paper, 2021
4. Shiwani Gupta "Apply Machine Learning for Healthcare to enhance performance and identify informative features", IEEE, Research Paper, 2021
5. Satyajit Pattnaik "Easy steps to deploy Machine Learning (ML) models in AWS EC2 Instance", 2021
6. CampusX "How To Deploy A Machine Learning Model On AWS EC2 | AUG 2021 Updated | ML Model To Flask Website", 2021
7. Raihan M, Mondal S, More A, Sagor M O F, Sikder G, Majumder M A & Ghosh K (2016, December). Smartphone based ischemic heart disease (heart attack) risk prediction using clinical data and data mining approaches, a prototype design. In 2016 19th International Conference on Computer and Information Technology (ICCIT) (pp. 299-303). IEEE.
8. Dangare C S & Apte S S (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44-8.
9. Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8