Web Application for Cardiovascular Diagnosis Using Data Science

M. Imrankhan[1] , S. Deivanai[2] , G. Lakshmi Priya[3], B. Vinothini[4]
[1] *Assistant Professor, Dept. of CSE, Manakula Vinayagar Institute of Technology, Puducherry*
[2,3,4] *Third Year Students, Dept. of CSE, Manakula Vinayagar Institute of Technology, Puducherry*

***Abstract*—Cardiovascular disease (CVD) is a big reason for morbidity and mortality in the current living style. Identification of Cardiovascular disease is an important but a complex task that needs to be performed very minutely, efficiently and the correct automation would be very desirable. With the help of data science tools and machine learning algorithms, doctors can detect and track common conditions, like cancer, cardiac or respiratory diseases. Data Science technology can also detect the slightest changes in the patient's health indicators and predict possible disorders. An automated system in medical diagnosis would enhance medical care and it can also reduce costs. In this study, we have designed a system that can efficiently discover the rules using dataset of major health factors from patients to predict the risk level of patients based on the given parameter about their health. The performance of the system is evaluated in terms of model prediction accuracy and the results show that the system has great potential in predicting the cardiovascular disease risk level more accurately.**

***Keywords*— Cardiovascular disease, Data Science, Machine Learning, Random forest algorithm, Prediction and Accuracy.**

I. INTRODUCTION

A. Overview

Data science combines math and statistics, specialized programming, advanced analytics, Artificial intelligence(AI), and Machine learning(ML) with specific subject matter expertise to uncover actionable insights hidden in an organization's data. These insights can be used to guide decision making and strategic planning. Data science uses complex machine learning algorithms to build predictive models. It Attempts to withdraw hidden pattern and trends from huge databases. It extracts Meaningful information from database which helps for future prediction. Healthcare companies are using data science to build sophisticated medical instruments to detect and cure diseases. Data Science helps in advancing healthcare facilities and processes. It helps boost productivity in diagnosis and treatment and enhances the workflow of healthcare systems.

B. Problem Definition

CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. Clinical decisions are often made based on doctor's experience and intuition rather and much cost that affect the medical services quality. Using analytic tools and data modeling can help in enhancing clinical decisions. Thus, the goal here is to build a web application to help doctors in diagnosing cardiovascular diseases.

C. Thesis Motivation

The main motivation of our thesis is to identify and cure the cardiovascular illness at the beginning stage itself, in order to reduce the morbidity and mortality rate. To produce a low cost and time-efficient tool to diagnose cardiovascular illness. It will create an awareness of a healthy lifestyle. It is a device to be aware of cardiovascular health. The main motto of our thesis is to produce exactly correct predicted values without any errors.

D. Research Objective

The main objective of this study is to take a step forward to reduce the mortality rate of patients with cardiovascular disease by providing all the essential precautions to diagnose. The proposed system diagnoses cardiovascular diseases by providing an effective and

efficient Web app using Data Science Engineering. The objective of this study is to make the cardiovascular diagnosis process easy, less time-taking, and secure prediction. A web application for cardiovascular diagnostic systems to eliminate false predictions that can happen in traditional physician prediction schemes.

II. SYSTEM STUDY

A. Existing System

The patient provides the input details for this system. The cardiac illness is then assessed using machine learning algorithms based on the user inputs. The generated findings are now compared to those of current models in the same domain and found to be superior. Patterns are discovered using NN, DT, Support Vector Machines SVM, and Naive Bayes on data from heart disease patients obtained at the UCI laboratory. With these algorithms, the performance and accuracy of the outcomes are compared. In comparison to other current methods, the suggested hybrid method produces results of 87 percent for F-measure.

B. Drawbacks

- Prediction of cardiovascular disease results is not accurate.
- There is no proper Data mining techniques are used.
- Cannot handle enormous datasets for patient records.
- There is no proper frontend tool to use the system
- Web app is difficult to use in the real time.

C. Existing System

In this project, python, seaborn, pyplot and pandas operations are to classify Cardiovascular disease using data from the UCI repository after reviewing the findings from existing approaches. The machine learning process begins with data pre-processing, then moves on to feature selection based on data cleansing, categorization, and evaluation of modelling performance. To improve the accuracy of the outcome, a Random Forest technique is applied. For creating the user friendly web app, here we are used Streamlit web framework with python for implementation of best cardiovascular diagnosis web app. Further we have created a QR code for opening the web app by importing the module pyqrcode and pypng using python code. So, anyone can use the web app at anytime at anywhere by simply scanning our QR code.

D. Advantages

- Increased accuracy of 100% for effective cardiovascular disease diagnosis.
- Reduce the time complexity of doctors and cost effective for patients.
- Well-designed web application for prediction with probability rate.
- QR code implementation for opening the Web app.
- Best user experience web app which can be used in all the devices.

E. Feasibility Study

The use case we have been taken is a serious problem in the field of healthcare. The system we have discovered will create an evolution to stop the cardiovascular diagnosis problem. Further, we implemented the web app using the QR code scanning method. Anyone can use the cardiovascular diagnosis web app anytime and anywhere by simply scanning the implemented QR code. The system we designed will produce the results with 100% accuracy that was deployed with the Random Forest ML model.

III. DATASETS

A. Collection of datasets

The dataset is collected from Kaggle, a subsidiary of Google LLC, is an online community of data Scientists and machine learning practitioners. Kaggle allows users to find and Publish data sets, explore and build models in a web-based data-science Environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. The UCI Machine Learning Repository is a collection of databases, Domain theories, and data generators that are used by the machine learning Community for the empirical analysis of machine learning algorithms. This is multivariate type of dataset which means providing or involving a variety of separate mathematical or statistical variables, multivariate numerical data analysis. It is composed of 14 attributes which are age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, oldpeak — ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels and Thalassemia. This database includes 76 attributes, but all published studies relate to the use of a subset of 14 of them. The Cleveland database is the only one used by ML researchers to date. One of the major

tasks on this dataset is to predict based on the given attributes of a patient that whether that particular person has a heart disease or not and other is the experimental task to diagnose and find out various insights from this dataset which could help in understanding the problem more.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1020	59	1	1	140	221	0	1	164	1	0.0	2	0	2	1
1021	60	1	0	125	258	0	0	141	1	2.8	1	1	3	0
1022	47	1	0	110	275	0	0	118	1	1.0	1	1	2	0
1023	50	0	0	110	254	0	0	159	0	0.0	2	0	2	1
1024	54	1	0	120	188	0	1	113	0	1.4	1	1	3	0

1025 rows × 14 columns

Fig. 1 Overview of Datasets

B. Dataset Attributes

- age: The person's age in years
- sex: The person's sex
  - 1 = male
  - 0 = female
- cp: chest pain type
  - Value 0: asymptomatic
  - Value 1: atypical angina
  - Value 2: non-anginal pain
  - Value 3: typical angina
- trestbps: The person's resting blood pressure (mm Hg on admission to the hospital)
- chol: The person's cholesterol measurement in mg/dl
- fbs: The person's fasting blood sugar (> 120 mg/dl)
  - 1 = true
  - 0 = false
- restecg: resting electrocardiographic results
  - Value 0: showing probable or definite left ventricular hypertrophy
  - Value 1: normal
  - Value 2: having ST-T wave abnormality
- thalach: The person's maximum heart rate achieved
- exang: Exercise induced angina
  - 1 = yes
  - 0 = no
- oldpeak: ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot.)
- slope: the slope of the peak exercise ST segment
  - 0: downsloping
  - 1: flat
  - 2: upsloping
- ca: The number of major vessels (0–3)
- thal: A blood disorder called thalassemia Value 0: NULL (dropped from the dataset previously
  - Value 1: fixed defect (no blood flow in some part of the heart)
  - Value 2: normal blood flow
  - Value 3: reversible defect (a blood flow is observed but it is not normal)
- target: Heart disease
  - 1 = no
  - 0= yes

C. Data Preprocessing

Data preprocessing transforms the data into a format that is more easily and effectively processed in data mining, machine learning and other data science tasks. The techniques are
generally used at the earliest stages of the machine learning and AI development pipeline to ensure accurate results.

There are several different tools and methods used for pre-processing data, including the following:

- Sampling, which selects a representative subset from a large population of data transformation, which manipulates raw data to produce a single input.
- Denoising, which removes noise from data.
- Imputation, which synthesizes statistically relevant data for missing values.
- Normalization, which organizes data for more efficient access.
- Feature extraction, which pulls out a relevant feature subset that is significant in a particular context.

These tools and methods can be used on a variety of data sources, including data stored in files
or databases and streaming data.

D. Attributes Correlation

Correlation indicates how the features are related to each other or to the target variable. The correlation may be positive (increase in one value of the feature increases the value of the
target variable) or negative (increase in one value of the feature decreases the value of the target variable) Heatmap makes it easy to classify the features are most

relevant to the target variable, and we will plot the associated features of the heatmap using the seaborn library. Correlation shows whether the characteristics are related to each other or to the target variable. Correlation can be positive (increase in one value, the value of the objective variable increases) or negative (increase in one value, the value of the target variable decreased). From this heatmap we can observe that the 'cp' chest pain is highly related to the target variable. Compared to the relation between the other two variables we can say that chest pain contributes the most in prediction of presences of a heart disease. A medical emergency is a heart attack. A cardiac occurs usually when blood clot blocks blood flow to the cardiac. Tissues lose oxygen without blood and die causing chest pain.

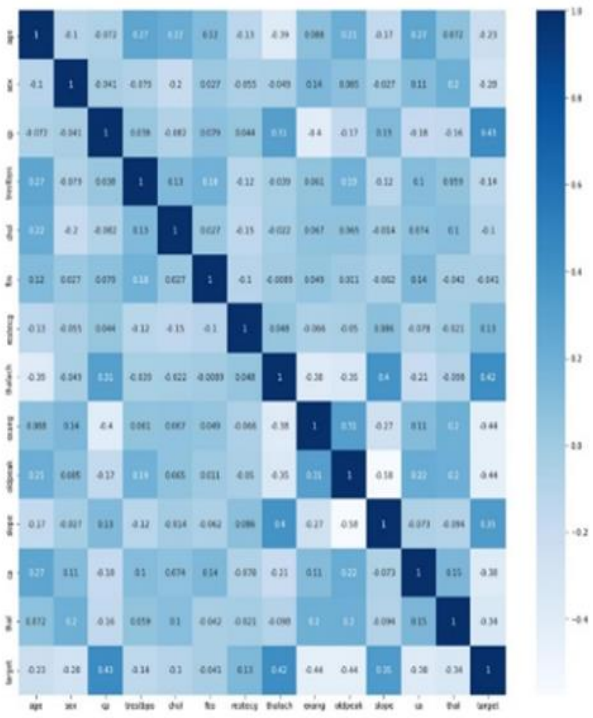


Fig. 2 Attributes correlation map

IV. EXPLORATORY DATA ANALYSIS

A. *EDA Overview*

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions. The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables. Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals.

B. *EDA Tools*

Some of the most common data science tools used to create an EDA include:

- Python: An interpreted, object-oriented programming language with dynamic semantics. Its high-level, built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for rapid application development, as well as for use as a scripting or glue language to connect existing components together. Python and EDA can be used together to identify missing values in a data set, which is important so you can decide how to handle missing values for machine learning.
- R: An open-source programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians in data science in developing statistical observations and data analysis.

C. *Data Visualization*

Data visualization techniques are the most important part of Data Science, There won't be any doubt about it. And even in the Data Analytics space as well the Data visualization doing a major role. We will discuss this in detail with help of Python packages and how it helps during the Data Science process flow. This is a very interesting topic for every Data Scientist and Data Analyst.

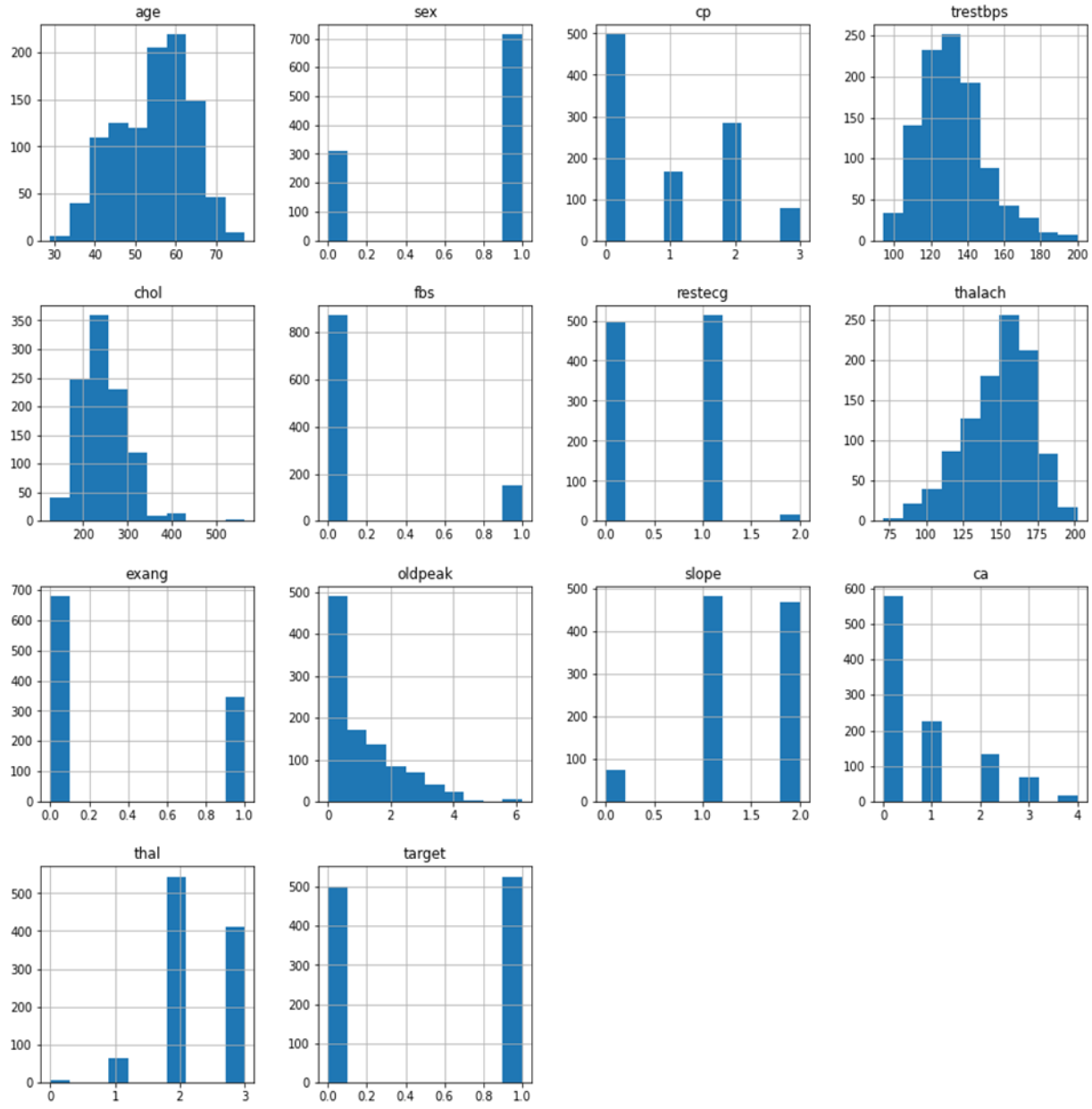


Fig. 3 Attributes Visulization

V. DATA MODELLING

A. Usage of ML algorithm

Overshadowing other data science aspects, machine learning (ML) and artificial intelligence (AI) have dominated the industry nowadays in the following ways:

- Machine learning analyzes and examines large chunks of data automatically.
- It automates the data analysis process and makes predictions in real-time without any human involvement.
- You can further build and train the data model to make real-time predictions. This point is where you use machine learning algorithms in the data science lifecycle.

Model training depends on both the quality of the training data and the choice of the machine learning algorithm. An ML algorithm is selected based on end-user needs. Once the right machine learning algorithm is selected, the training data set is divided into two parts for training and testing in the 7:3 ratio. This is done to determine the bias and variance of the ML model. As a result of model

training, you will achieve a working model that can be further validated, tested, and deployed.

B. Logistic Regression

Logistic Regression is used to estimate discrete values (usually binary values like 0/1) from a set of independent variables. It helps predict the probability of an event by fitting data to a logit function. It is also called logit regression. These methods listed below are often used to help improve logistic regression models:

- Include interaction terms.
- Eliminate features.
- Regularize techniques.
- Use a non-linear model.

C. SVM

SVM algorithm is a method of a classification algorithm in which you plot raw data as points in an n-dimensional space (where n is the number of features you have). The value of each feature is then tied to a particular coordinate, making it easy to classify the data. Lines called classifiers can be used to split the data and plot them on a graph.

D. Decision Tree

Decision Tree algorithm in machine learning is one of the most popular algorithm in use today; this is a supervised learning algorithm that is used for classifying problems. It works well in classifying both categorical and continuous dependent variables. This algorithm divides the population into two or more homogeneous sets based on the most significant attributes/ independent variables.

E. Random Forest

A collective of decision trees is called a Random Forest. To classify a new object based on its attributes, each tree is classified, and the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).Each tree is planted & grown as follows:

- If the number of cases in the training set is N, then a sample of N cases is taken at random. This sample will be the training set for growing the tree.
- If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M, and the best split on this m is used to split the node. The value of m is held constant during this process.
- Each tree is grown to the most substantial extent possible. There is no pruning.

F. KNN

This algorithm can be applied to both classification and regression problems. Apparently, within the Data Science industry, it's more widely used to solve classification problems. It's a simple algorithm that stores all available cases and classifies any new cases by taking a majority vote of its k neighbours. The case is then assigned to the class with which it has the most in common. A distance function performs this measurement. KNN can be easily understood by comparing it to real life. Things to consider before selecting K Nearest Neighbours Algorithm:

- KNN is computationally expensive.
- Variables should be normalized, or else higher range variables can bias the algorithm.
- Data still needs to be pre-processed.

G. Adaboost Classifier

AdaBoost algorithm, short for Adaptive Boosting, is a Boosting technique used as an Ensemble Method in Machine Learning. It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances. Boosting is used to reduce bias as well as variance for supervised learning. It works on the principle of learners growing sequentially. Except for the first, each subsequent learner is grown from previously grown learners. In simple words, weak learners are converted into strong ones.

H. Gradient Boosting

Gradient Boosting is a functional gradient algorithm that repeatedly selects a function that leads in the direction of a weak hypothesis or negative gradient so that it can minimize a loss function. Gradient boosting classifier combines several weak learning models to produce a powerful predicting model. Gradient boosting classifier requires these steps:

- Fit the model.
- Adapt the model's Hyperparameters and Parameters.
- Make forecasts.
- Interpret the findings.

VI. DATA EVALUATION

A. Confusion matrix

A confusion matrix is a tabular summary of the number of correct and incorrect predictions made by a classifier. It is used to measure the performance of a classification

model. It can be used to evaluate the performance of a classification model through the calculation of performance metrics like accuracy, precision, recall, and F1-score.

The following 4 are the basic terminology which will help us in determining the metrics we are looking for.

- True Positives (TP): when the actual value is Positive and predicted is also Positive.
- True negatives (TN): when the actual value is Negative and prediction is also Negative.
- False positives (FP): When the actual is negative but prediction is Positive. Also known as the Type 1 error.
- False negatives (FN): When the actual is Positive but the prediction is Negative .Also known as the Type 2 error.

B. Accuracy

It is a measure of correctness that is achieved in true prediction. In simple words, it tells us how many predictions are actually positive out of all the total positive predicted. Accuracy is a valid choice of evaluation for classification problems which are well balanced and not skewed or there is no class imbalance.

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

C. Precision

It is a measure of correctness that is achieved in true prediction. In simple words, it tells us how many predictions are actually positive out of all the total positive predicted. Precision is defined as the ratio of the total number of correctly classified positive classes divided by the total number of predicted positive classes.

Precision = TP / (TP + FP)

D. Recall

It is a measure of actual observations which are predicted correctly, i.e. how many observations of positive class are actually predicted as positive. It is also known as Sensitivity. Recall is a valid choice of evaluation metric when we want to capture as many positives as possible. Recall is defined as the ratio of the total number of correctly classified positive classes divide by the total number of positive classes.

Recall = TP / (TP + FN)

E. F1 – Score

The F1 score is a number between 0 and 1 and is the harmonic mean of precision and recall. We use harmonic mean because it is not sensitive to extremely large values, unlike simple averages. F1 score sort of maintains a balance between the precision and recall for your classifier. If your precision is low, the F1 is low and if the recall is low again your F1 score is low.

F1-Score = 2 * ((Recall * Precision)/(Recall+Precision))

F. Support

Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn't change between models but instead diagnoses the evaluation process.

VII. MODEL DEPLOYMENT

A. Performance analysis

TABLE I Performance analysis of various ML model

ML MODEL	TRAINING ACCURACY%	TESTING ACCURACY%
Logistics Regression	84.79	86.68
Support Vector Machine	69.17	75.0
Decision Tree	100.0	100.0
Random Forest	100.0	100.0
K Nearest Neighbor	71.68	73.05
Gradient Boosting	91.49	89.61
Ada Boosting	98.60	97.72

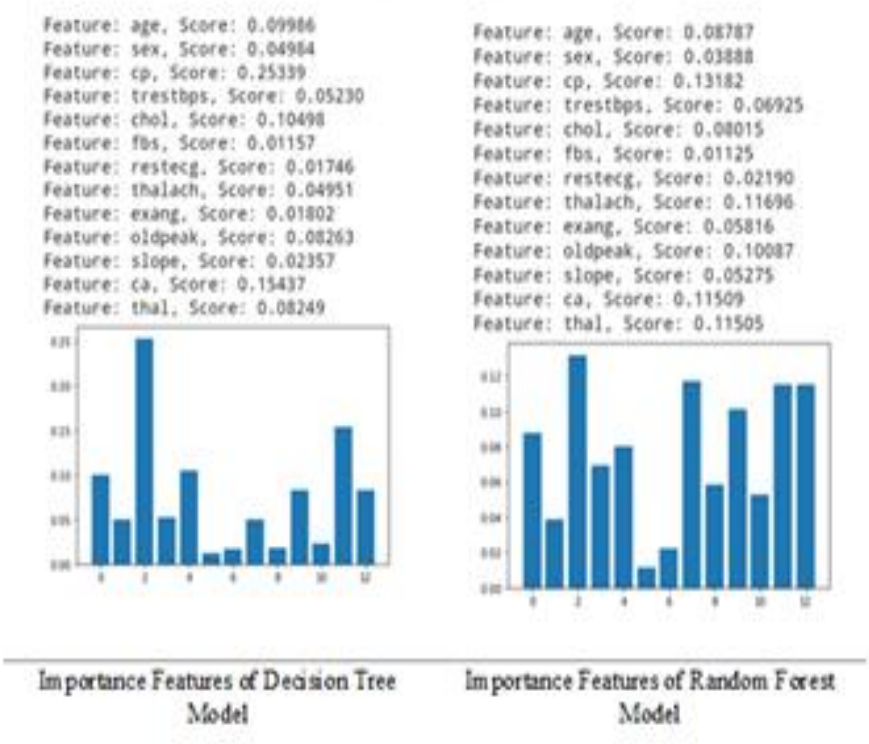
Here, we have been trained and tested the datasets with different ML algorithm model for building the further prediction model for developing the web application which can be used for future clinical data analysis. One of the most important key feature to deploy a prediction model is to select the best ML model which will suits for the given Use case's datasets relationship properly and efficiently. From the Table, we can see that the most common error in developing ml model that is Overfitting and Underfitting is occurred in some cases of ml algorithm that means those algorithms are not fitted for the given datasets. Whereas Overfitting means that your model makes not accurate predictions. In this case, train error is very small and val/test error is large. Underfitting

means that your model makes accurate, but initially incorrect predictions. In this case, train error is large and val/test error is large too. So, we reject those algorithms. Finally, we can noticed that 100% Accuracy is produced for two algorithms namely Decision Tree and Random Forest. That is its shows that given datasets is perfectly suited for Decision Tree and Random Forest.

B. Performance measures

The critical difference between the random forest algorithm and decision tree is that decision trees are graphs that illustrate all possible outcomes of a decision using a branching approach. In contrast, the random forest algorithm output are a set of decision trees that work according to the output. Random forest algorithm is the Ensemble model. Ensembles methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly. Random Forest is suitable for situations when we have a large dataset, and interpretability is not a major concern. So , we choosing Random Forest ML model as our final Prediction model for developing web application.

TABLE 2
Decision tree vs Random forest



VIII. SYSTEM IMPLEMENTATION

A. Flow Chart

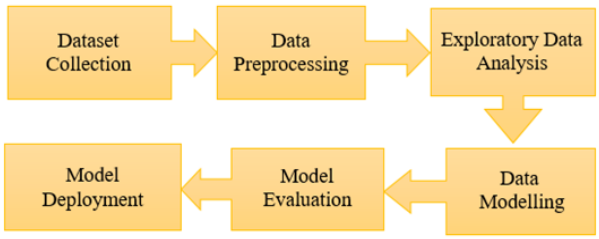


Fig. 4 System flow chart

- Data collection - In this step we need collect the relevant data to get the desired output.
- Data Preprocessing - This stage helps us gain a better understanding of the data and prepares it for further evaluation.
- Exploratory data analysis - This step helps us rule out unwanted models from the set of wide variety of models that we have in hand. This is the core process

of a data science life cycle, where the correct model type is selected.

- Data modelling - At this stage we need to split the cleaned dataset from the previous step into train and test sets. We will train the various ML model.
- Model evaluation - The model build and evaluation phase is where you do the actual modelling of the data.
- Model deployment - This step creates the delivery mechanism you need to get the model out to the users or to another system

B. System Architecture

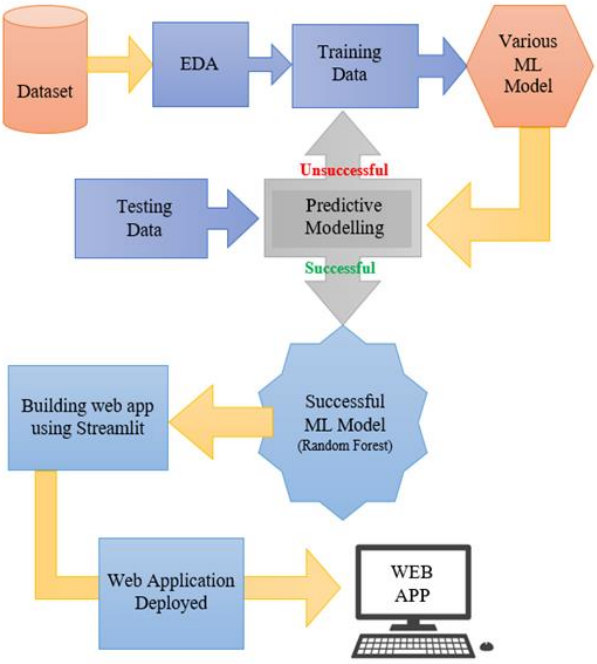


Fig. 5 System Architecture diagram

C. Methodology

For the proposed study dataset was taken from both Kaggle and UCI ml repository site. Then it was downloaded in excel file using comma separated format. Data has processed by python programming using Google Colab notebook. The data set contains 1024 sample instances as shown in Fig. 3.1. The dataset contains 14 clinical features as shown in chapter-3. Different types of python libraries such as pandas, Sklearn, NumPy, Matplotlib are used for processing the algorithms. Using explorative data analysis technique data was analysed in Google Colab notebook. 10-fold cross validation technique is used for spitting the data set into training and testing data in the 7:3 ratio. Data Modelling had been performed with various ml algorithms as shown in Table-1. Then the research model was evaluated using supervised learning techniques such as random forest and decision trees. Further we have chosen the Random forest ml model for implementation of the future prediction model. This ml prediction model has developed as the Web application for diagnosis. Here, we have used Streamlit Web framework with python programming language for implementing the Web application. Streamlit is an open source app framework in Python language. It helps us create web apps for data science and machine learning in a short time. A faster way to build and share data apps. Streamlit turns data scripts into shareable web apps in minutes. We will run the code using command prompt in python virtual environment Now the web app will run locally at web browser using Local host.

For further implementation, the web app is deployed in the Streamlit cloud using GitHub repository. Finally, web app URL is generated using Streamlit. QR code is generated for the web application URL using python code by importing the pyqrcode and pypng module. By scanning the QR code anyone can open web application in the web browser then user can enter the clinical attributes data which will produce the diagnosis result precisely.

Fig. 6 QR Code of our implemented system

IX. CONCLUSION

Identifying the processing of raw healthcare data of cardiovascular information will help in the long term saving of human lives and early detection of abnormalities in cardiovascular conditions. In this project, we proposed a machine learning-based strategy using data science domain for predicting cardiovascular

illness, and the findings demonstrated a high accuracy threshold for offering a superior estimation result. Cardiovascular disease prediction is challenging and very important in the medical field. However, the mortality rate can be drastically controlled if the disease is detected at the early stages and preventative measures are adopted as soon as possible. Further extension of this study is highly desirable to direct the investigations to real-world datasets instead of just theoretical approaches and simulations. The proposed Random Forest model proved to be quite accurate in the prediction of cardiovascular disease. We obtained the Accuracy value of 100% for prediction. ML Techniques are used to extract information from the input dataset using data science. Finally, the web application is developed for the prediction model at low cost with high efficiency.

ACKNOWLEDGMENT

We express our sincere and heartfelt thanks to our Chairman and Managing Director, Vice Chairman and our Secretary for their support and encouragement. It gives me great ecstasy of pleasure to convey my deep and sincere thanks to our principal, for giving me constant motivation to succeed in my goal. With profoundness we would like to express my sincere thanks to our beloved Head of the Department, and Associate/Assistant Professors of Department of Computer Science and Engineering, for their kindness in extending the infrastructural facilities to carry out our project work successfully. We extend our sincere gratitude and heartfelt thanks to this Journal for publishing our paperwork successful. Hope this Journal will encourage us to publish more research works on emerging technologies in future.

REFERENCE

[1] Chen, A. H., Huang, S. Y., Hong, P. S., Cheng, C. H., & Lin, E. J. (2011,September). HDPS: “heart disease prediction system”. In 2011 Computing in Cardiology (pp. 557-560). IEEE.

[2] Shetty, Deeraj, Kishor Rit, Sohail Shaikh, and Nikita Patil. "Diabetes diseas eprediction using data mining."In 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1-5.IEEE, 2017.

[3] Rajesh , T Maneesha, Shaik Hafeez, Hari Krishna“Prediction of Heart Disease Using Machine Learning Algorithms“May 2018International Journal of Engineering & Technology 7(2):363-366DOI: 10.14419/ijet. v7i2.32.15714 North-Holland/American Elsevier) p 517.

[4] J. Krishnan Santana; S. Geetha “Prediction of Heart Disease Using Machine Learning Algorithms”. 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)Publisher: IEEE

[5] Rajdhan Apurb, Agarwal Avi, Sai Milan, Ravi Dundigalla, Ghuli Poonam.”Heart Disease Prediction using Machine Learning” INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY

[6] Singh, A., & Kumar, R. (2020). “Heart Disease Prediction Using Machine Learning Algorithms”. 2020 International Conference on Electrical and Electronics Engineering (ICE3). doi:10.1109/ice348803.2020.9122958

[7] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). “Effective Heart Disease Prediction using Hybrid Machine Learning Techniques”. IEEE Access, 1–1. doi:10.1109/access.2019.2923707

[8] Rairikar, A., Kulkarni, V., Sabale, V., Kale, H., & Lamgunde, A. (2017,June). “Heart disease prediction using data mining techniques”. In 2017 International Conference on Intelligent Computing and Control(I2C2) (pp. 1-8). IEEE

[9] Al Essa, Ali Radhi, and Christian Bach. "Data Mining and Warehousing." American Society for Engineering Education (ASEE Zone 1) Journal (2014).

[10] National Health Council, ‘Heart HealthScreenings’,2017. [Online]Available: http://www.heart.org/HEARTORG/Conditions/HeartHealthScreenings_UCM_4 28687_Article.jsp#. WnsOAeeYPIV

[11] M. Durairaj and V. Revathi, ‘‘Prediction of heart disease using back propagation MLP algorithm,’’ Int. J. Sci. Technol. Res., vol. 4, no. 8, pp. 235–239,2015.

[12] M. Gandhi and S. N. Singh, ‘‘Predictions in heart disease using techniques of data mining,’’ in Proc. Int. Conf. Futuristic Trends Comput. Anal. Knowl.Manage. (ABLAZE), Feb. 2015, pp. 520–525.

[13] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, ‘‘Prediction of heart disease using machine learning,’’ in Proc. 2nd Int. Conf. Electron.,

Commun. Aerosp. Technol. (ICECA), Mar. 2018, pp. 1275–1278.

[14] B. S. S. Rathnayakc and G. U. Ganegoda, ‘‘heart diseases prediction with data mining and neural network techniques,’’ in Proc. 3rd Int. Conf. Converg. Technol. (I2CT), Apr. 2018, pp. 1–6.

[15] N. K. S. Banu and S. Swamy, ‘‘Prediction of heart disease at early stage using data mining and big data analytics: A survey,’’ in Proc. Int. Conf. Elect., Electron., Commun., Comput. Optim. Techn. (ICEECCOT), Dec. 2016, pp. 256–261.

[16] J. P. Kelwade and S. S. Salankar, ‘‘Radial basis function neural network for prediction of cardiac arrhythmias based on heart rate time series,’’in Proc. IEEE 1st Int. Conf. Control, Meas. Instrum. (CMI), Jan. 2016, pp. 454–458.