

Heart Disease Risk Predictor System using Data Mining Learning Techniques: Analysis

Ashutosh Kumar Singh
ABES Engineering College
Ghaziabad

Asmita Dixit
ABES Engineering College
Ghaziabad

Aatif Jamshed
ABES Engineering College
Ghaziabad

ABSTRACT

In this paper author throws the light on how the cast amount of data developed by healthcare firms is not utilized properly for making better decisions. The main objective of this research is developing such a model Intelligent Heart Disease Prediction System (IHDPS) with the help of three data mining modelling techniques, namely, Decision Trees prediction model, Naïve Bayes probability technique and Neural Network. The Intelligent Heart Disease Prediction System (IHDPS) enables better approach in finding and extricating hidden knowledge (patterns and relationships) associated with heart disease from previous heart disease repository. It enables answering complex queries for diagnosing heart disease and thus assisting in healthcare practitioners. It helps in making intelligent clinical decisions.

Keywords

Heart Disease, MVT, Django, Logic Regression

1. INTRODUCTION

A large amount of data is produced everyday but most of it is not used effectively as there are not much efficient tools to extricate knowledge from the databases. This may be thought of as treasure of hidden information that is mostly unhindered.[2]

In order to give better services in limited affordable costing, the healthcare industry is still struggling for a plan. Today, many hospitals manage healthcare data using healthcare information systems. These systems contain huge amount of data, used to extract hidden information for making intelligent medical diagnosis.

According to survey of WHO [2], 17 million total global deaths are due to [1] heart attacks and heart strokes. Bad clinical decisions can result in risky consequences which may not be accepted by the people. Its prediction is not easy as it requires high expertise and accuracy. The main aim of this project is to help the not so specialized doctors to take proper and better decisions about the risk of heart disease for a patient. The main objective of this research is to build Intelligent [2] Heart Disease Prediction System that gives diagnosis of heart disease using historical heart database. It aims at bringing together medicinal decision support with computer-based results of patient data, which helps to reduce medicinal faults and brings improvement in the overall outcome. Such type of work can bring together all the available data, as a basis on which development of rational assumption about the future could be done.

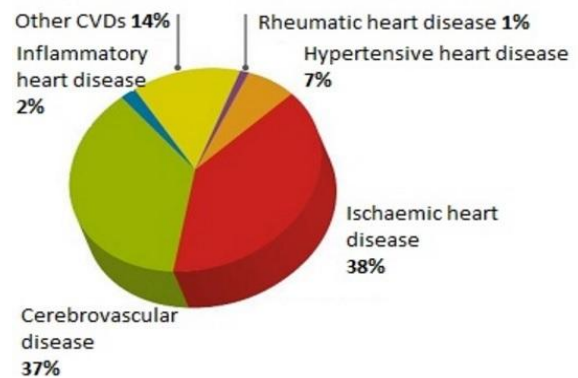


Fig. 1 Percentage of heart diseases

Diffrent type of Heart Disease

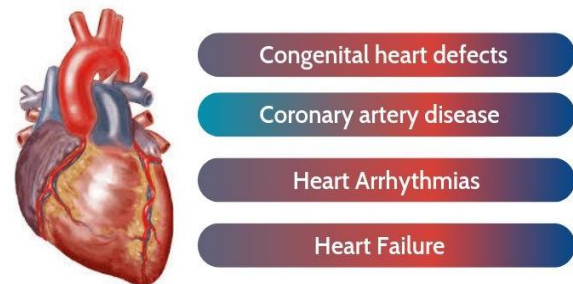


Fig .2 Types of heart diseases

2. RELATED WORK

Analysis of Data Mining Techniques for Heart Disease Risk Indicator

AUTHORS-MarjiaSultana,AfrinHaider and Mohammad Shorif Uddin

Heart disease is considered as one of the major cause of deaths throughout the world. It cannot be easily predicted by the medical practitioners as it requires high expertise and accuracy. This paper addresses the issue of prediction of [1] heart disease according to some input attributes of a person on the basis of data mining techniques.They have investigated the heart disease prediction through the Weka software using J48, SMO, Bayes, Net, KStar and Multilayer Perceptron.The performance of these data mining techniques is compared and measured by adding the results of predictive accuracy, ROC curve and AUC value using collected data and standard data. Based on performance factor SMO andBayes Net techniques show optimum performances than that of J48, KStar and Multilayer Peceptron techniques. [2]

Machine Learning Application to Predict the Risk of Coronary Artery Atherosclerosis

AUTHORS-SoodehNikan, FemidaGwadry-Sridhar, and Michael Bauer [2]

Coronary artery disease is the leading cause of death in the world. In this paper, they put forward an [1] algorithm based on the machine learning techniques to predict the risk of coronary artery atherosclerosis. A ridge expectation maximization imputation (REMI) [9] technique is proposed to find out and estimate the missing values in atherosclerosis databases. A conditional likelihood maximization method is used to remove irrelevant attributes and reduce the size of feature space and thus improve the speed of the learning. To evaluate the proposed algorithm, the UCI and [8] STULONG databases are used. The performance of heart disease prediction for two classification models is analysed and compared to previous work. The results showed improvement in percentage of risk prediction of the proposed method compared to other works. The effect of left out values in algorithm for prediction is also evaluated and the proposed REMI approach performs significantly better than conventional techniques.[3]

Analysis of Heart Disease Prediction using Various Machine Learning Techniques

Analysis of Heart Disease Prediction using Various machine Learning Techniques

AUTHORS-M. Marimuthu, S.Deivarani, Gayathri.R

In this paper, they try to concentrate on heart disease prediction. Using machine learning techniques, the heart disease can be predicted. The medical data such as Blood pressure, hypertension, diabetes, cigarette smoked per day and so on is taken as input and then these features are modelled for prediction. This model can then be used to predict future medical data. The algorithms like K-nearest neighbour, Naïve Bayes, support vector machine and decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the heart disease. [2]

3. METHODOLOGY

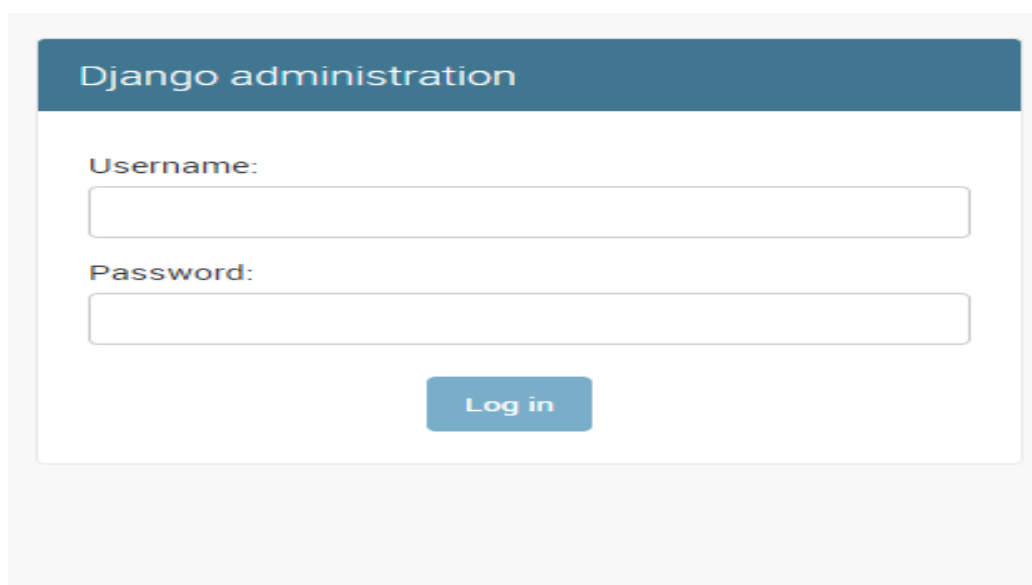
For this application firstly the user has to connect his phone to the internet. Then after login the user must register into the application using his basic personal details like username and email. Now to predict heart disease, user enters the values of various parameters on the basis of which risk factor will be calculated. After clicking predict button the result will be shown. If result is 1, user has risk of heart disease, if 0 then user does not have the risk. [3]

3.1 Implementation-

The implementation of this project is divided into two parts the Implementation of Prediction Engine and Implementation of Web Application. [8] To predict the risk of heart disease various attributes like cholesterol, blood, sugar, sex, etc. are used in prediction engine. The engine was developed in four increments. The libraries used in this are: [9]

- numpy: To work with arrays
- pandas: To work with csv files and dataframes
- matplotlib: To create charts using pyplot, using rcParams and colour using cm.rainbow
- warnings: To ignore all warnings due to past/past future depreciation of a feature
- train_test_split: To split the dataset into training and testing data
- StandardScaler: To scale all the features

“Jupyter Notebook” was used in first iteration whose objectives were to visualize the dataset, to find the accuracy in the prediction of the system, to create the general work flow of the prediction and to find any correlation between features. In the second increment the Jupyter Notebook was discarded and the code from the notebook was extracted, the visualizations were removed and the codes for prediction using different algorithms were extracted to create functions. The functionality was developed to show different types of scores for the trained models and to store and load trained models. Web API was introduced in third increment for predicting heart disease. This increment uses django for web framework and Forms for form validation[2]



The image shows a screenshot of the Django administration interface. At the top, there is a dark blue header with the text "Django administration" in white. Below the header, the page is white and contains a login form. The form has two input fields: "Username:" and "Password:". Below these fields is a blue button with the text "Log in" in white. The entire form is enclosed in a light gray border.

Fig.3 Django Administration

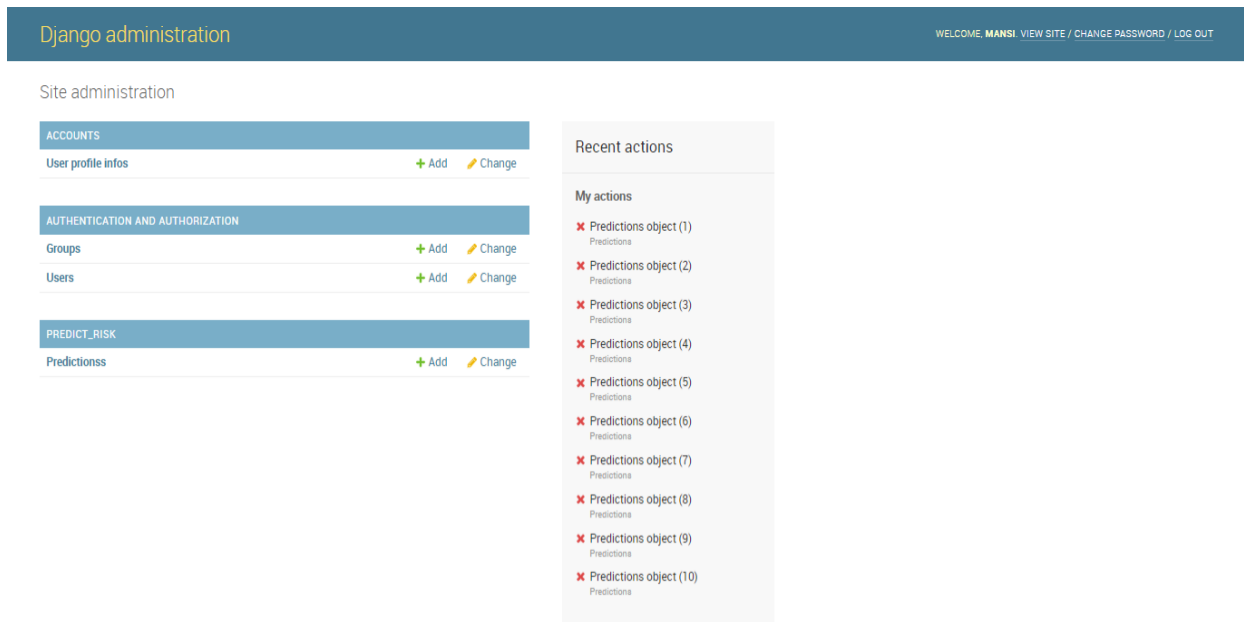


Fig.4 Django site administration

The web application has been developed using Django framework of python. It utilizes the [4] MVT (Modal View Template) architecture. The web application is implemented

through feature driven development and each activity of feature driven development is discussed with artefacts.

Table: 1 Performance Evaluation of MI Algorithms

ALGORITHMS	ACCURACY
LOGISTIC REGRESSION	83.88%
NAÏVE BAYES	80.73%
DECISION TREE	80.54%
SUPPORT VECTOR CLASSIFICATION	87.89%

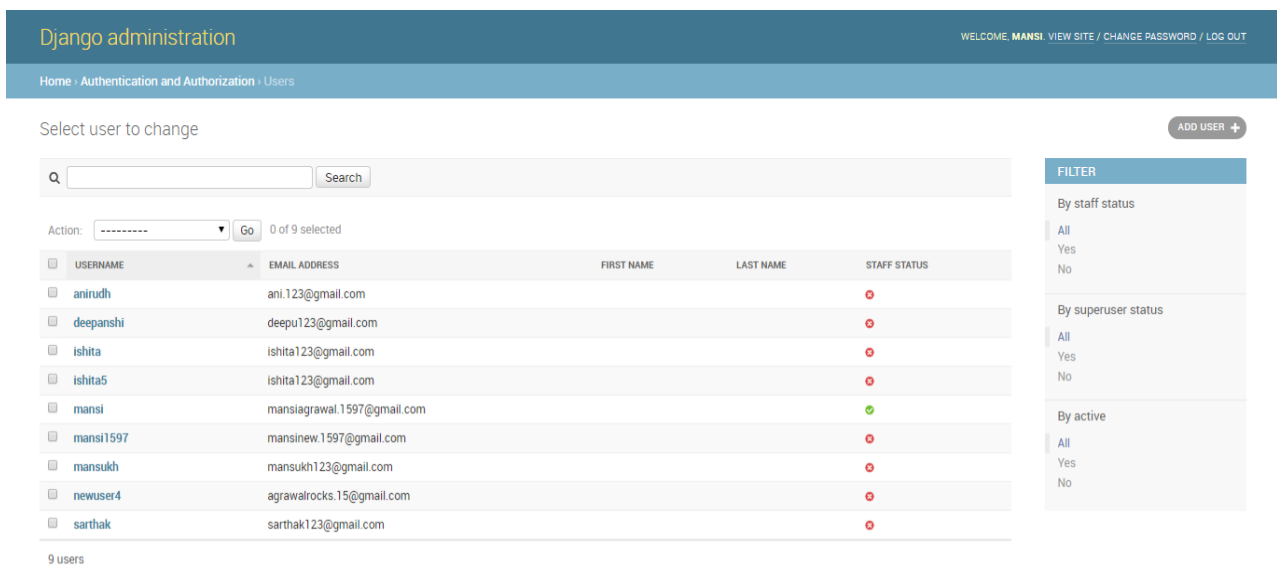


Fig.5 User Database

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num							
2	63	1	1	145	233	1	2	150	0	2.3	3	0	6	0							
3	67	1	4	160	286	0	2	108	1	1.5	2	3	3	1							
4	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1							
5	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0							
6	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0							
7	56	1	2	120	236	0	0	178	0	0.8	1	0	3	0							
8	62	0	4	140	268	0	2	160	0	3.6	3	2	3	1							
9	57	0	4	120	354	0	0	163	1	0.6	1	0	3	0							
10	63	1	4	130	254	0	2	147	0	1.4	2	1	7	1							
11	53	1	4	140	203	1	2	155	1	3.1	3	0	7	1							
12	57	1	4	140	192	0	0	148	0	0.4	2	0	6	0							
13	56	0	2	140	294	0	2	153	0	1.3	2	0	3	0							
14	56	1	3	130	256	1	2	142	1	0.6	2	1	6	1							
15	44	1	2	120	263	0	0	173	0	0	1	0	7	0							
16	52	1	3	172	199	1	0	162	0	0.5	1	0	7	0							
17	57	1	3	150	168	0	0	174	0	1.6	1	0	3	0							
18	48	1	2	110	229	0	0	168	0	1	3	0	7	1							
19	54	1	4	140	239	0	0	160	0	1.2	1	0	3	0							
20	48	0	3	130	275	0	0	139	0	0.2	1	0	3	0							
21	49	1	2	130	266	0	0	171	0	0.6	1	0	3	0							
22	64	1	1	110	211	0	2	144	1	1.8	2	0	3	0							
23	58	0	1	150	283	1	2	162	0	1	1	0	3	0							
24	58	1	2	120	284	0	2	160	0	1.8	2	0	3	1							
25	58	1	3	132	224	0	2	173	0	3.2	1	2	7	1							

Fig.6 Dataset

The UCI machine-learning repository has been used to take Cleveland Heart Disease Dataset(Lichman,2013). The

Repository contains 351 datasets maintained by University of California,Irvine.[1]

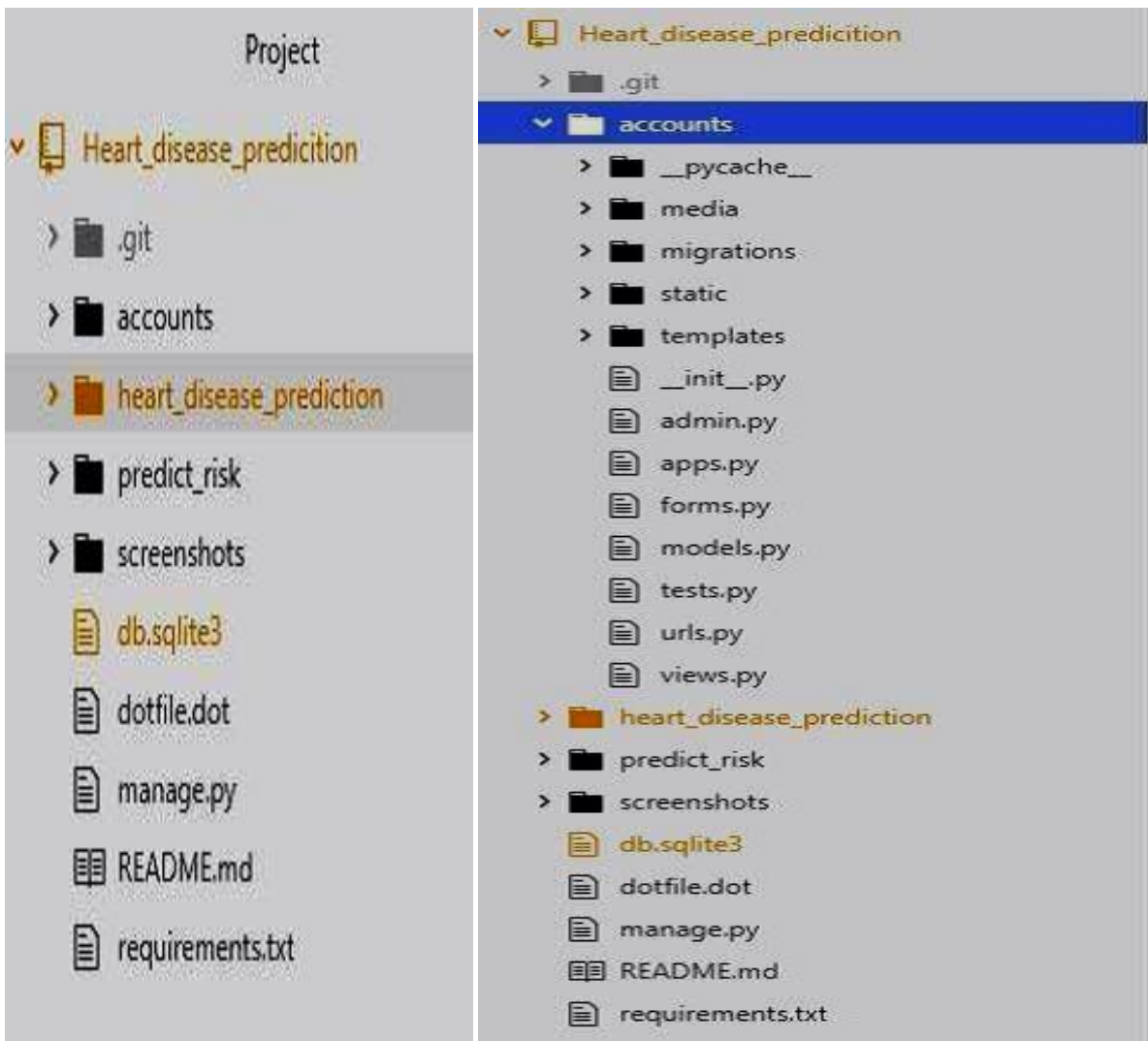


Fig:7 Project Architecture

3.2 Development of Overall Model-

Software requirement specification document was prepared for capturing the requirements during future driven development. ER Diagram and requirement specification document was designed. For the completion of the activity, a domain object model was prepared along with the overall application architecture.[4]

3.2.1 Functional Requirements-

- The system provides login for admin.
- The system should allow administrator to monitor and remove inappropriate datasets and code.

- The system allows users to create an account and login.
- The system allows the users to predict heart disease.
- The system allows the users to update their profile and password.[5]

3.3.1 Non-Functional Requirements-

- The website should provide values used during prediction to the user.
- The website should be responsive and consistent.

3.3 Architecture Diagram

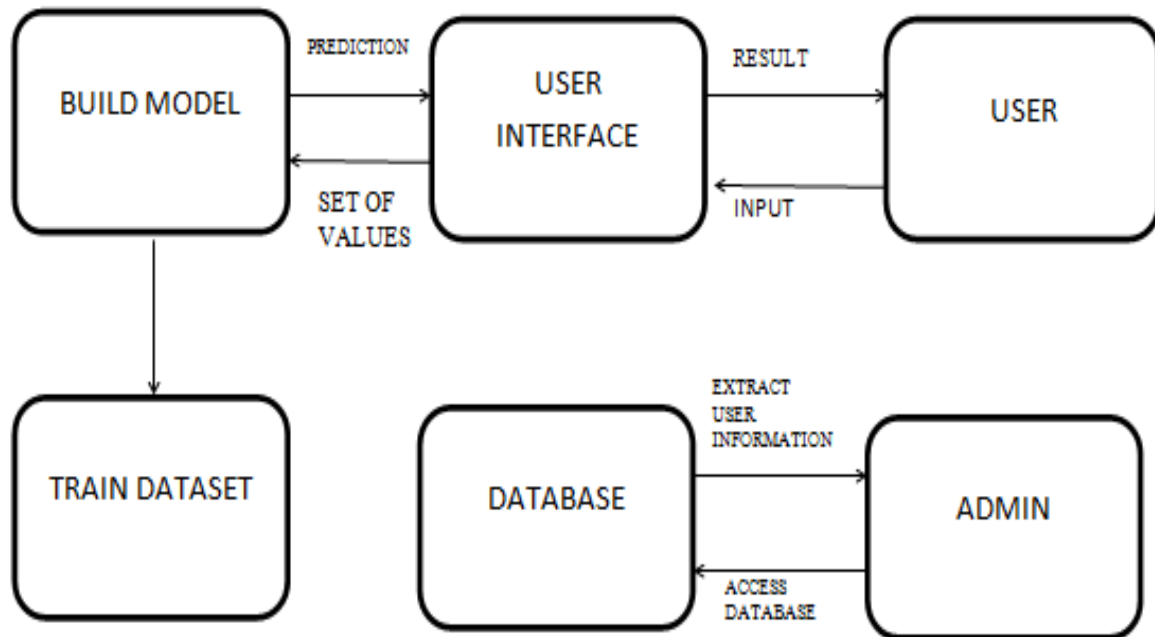


Fig.8 Architecture Methodology

Architecture-

The major component of Django project are models, views, templates along with urls.py, setting.py, admin.py, etc.

Models-

Models is a single, definitive data source which contains the essential field and behaviour of the data. In a database one model is one table and the field of the table is the attribute of the model. For this Django provides set of automatically-generated database application programming interfaces.

View-

View is a file containing python function which takes web requests and returns web responses. It is short form of view file. To link view function with URLs URLconf is used.

Template-

Django’s template is a text file which contains tags and variables. Logic of the template is controlled by tags. When the template is evaluated the result replaces the variables. Then Django works as a controller and check the availability of the resource in the URL. If the URL matches with any view, then Django returns it to the exact result of client related template as a quick response. [6]

4. TESTING AND RESULTS

The interface shows results as showed in images below. If the output of the corresponding result is 1 for more than two algorithms, then user has a “risk of heart disease”

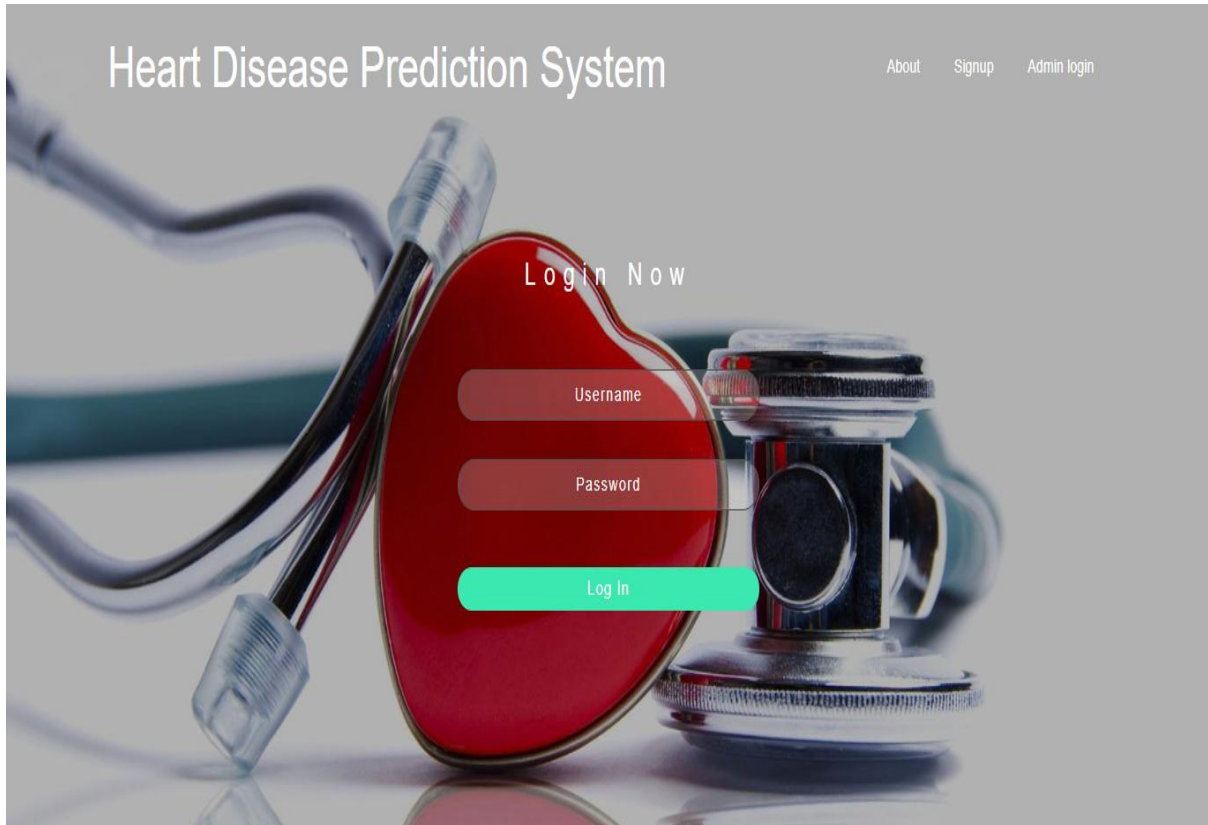


Fig.9 Login Page

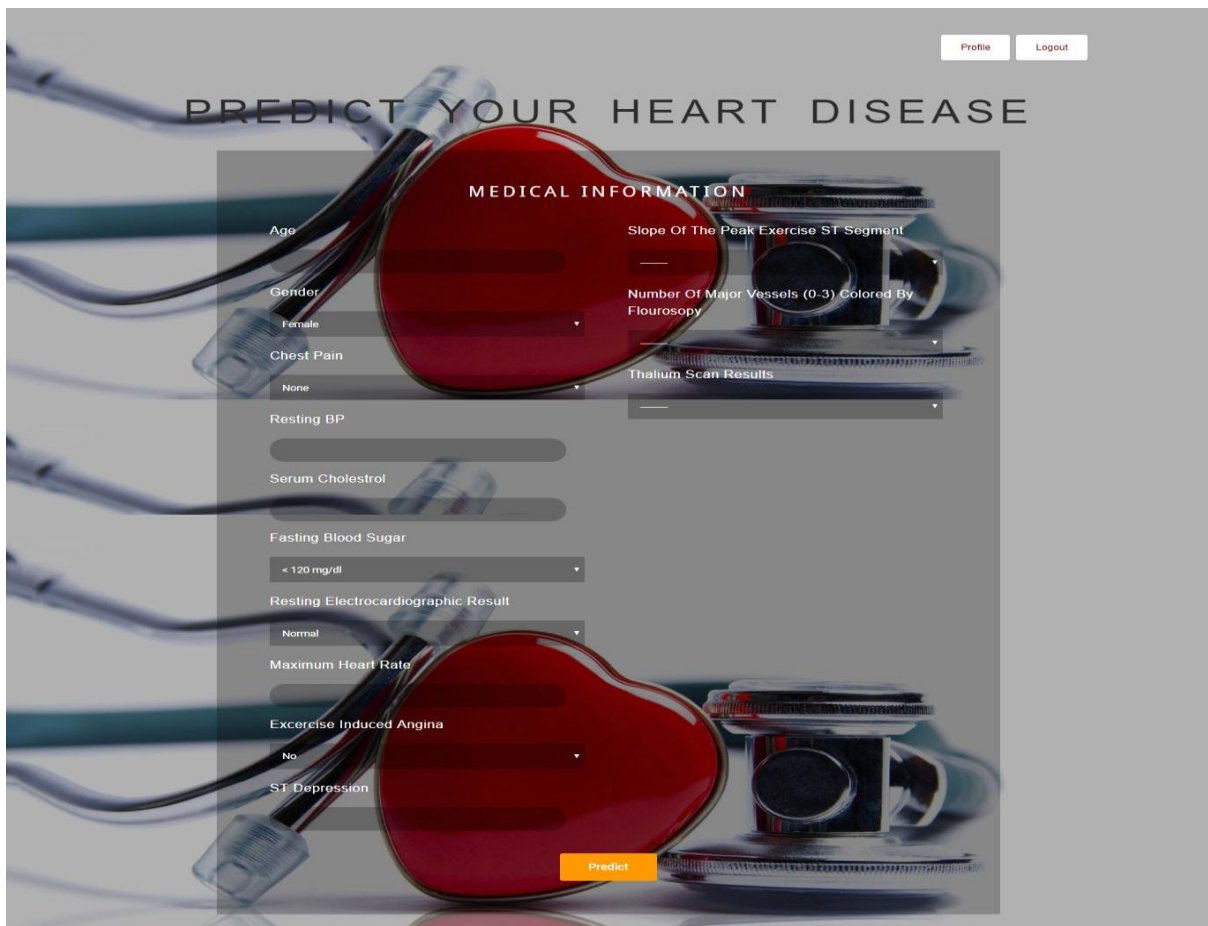


Fig.10 Medical Information Page

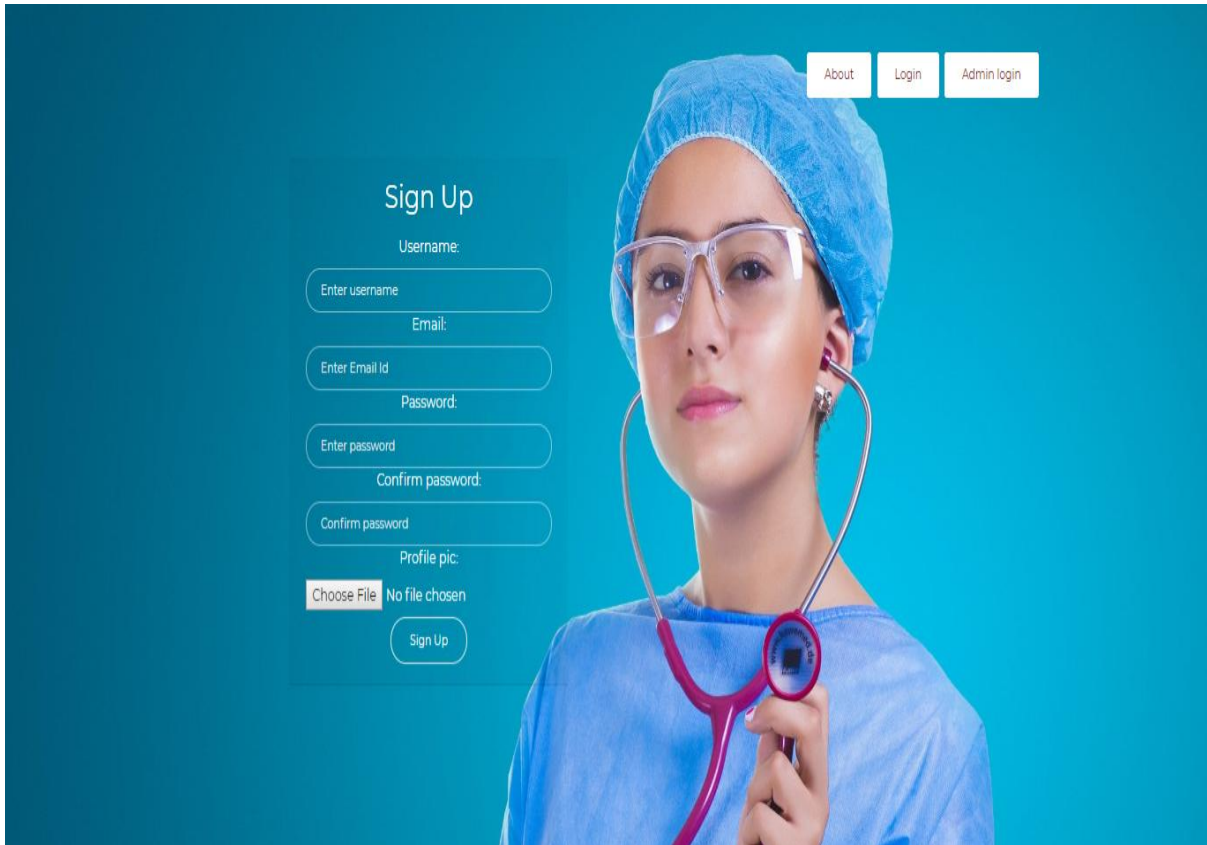


Fig.11 Sign up Page

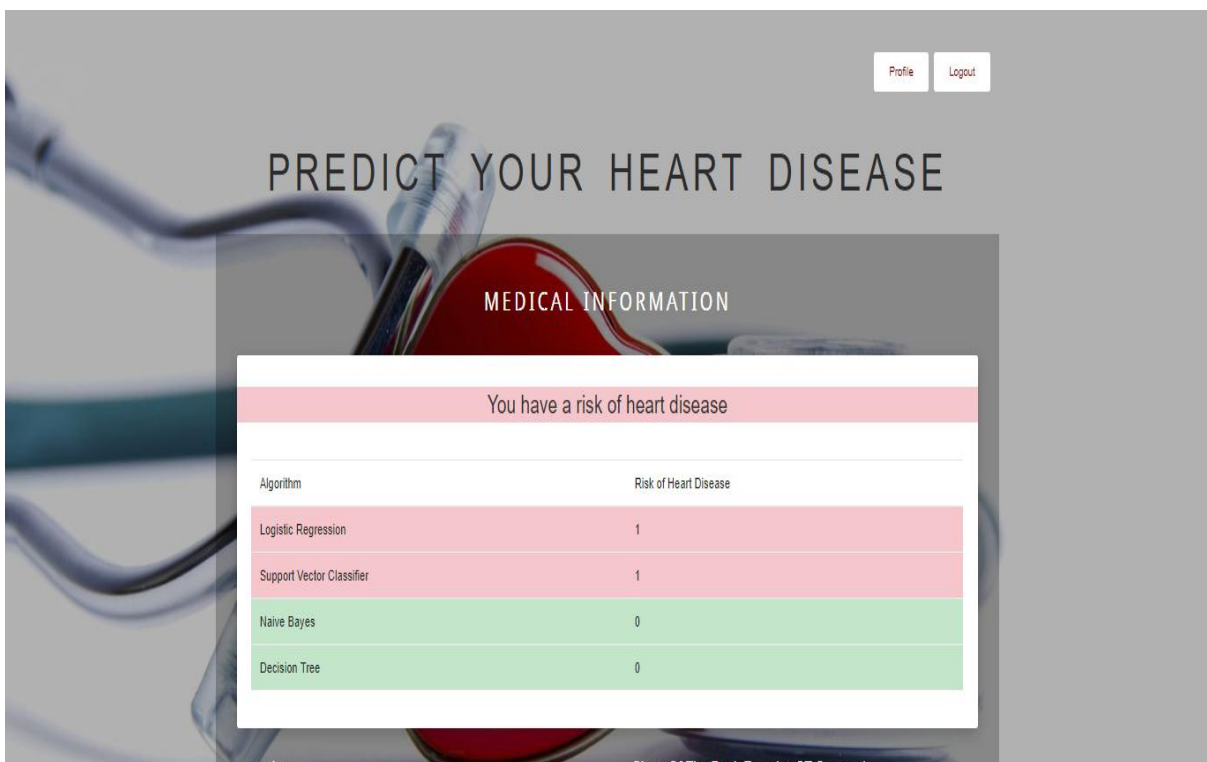


Fig.12 Prediction Results Page

The area under [7] ROC curve is a measure of how well a parameter can be used to distinguish between two diagnostic groups. In figure the AUC is higher for SVM followed by

Logistic Regression algorithm so we can say SVM performs rigorously slightly better than other logistics regressions algorithm.

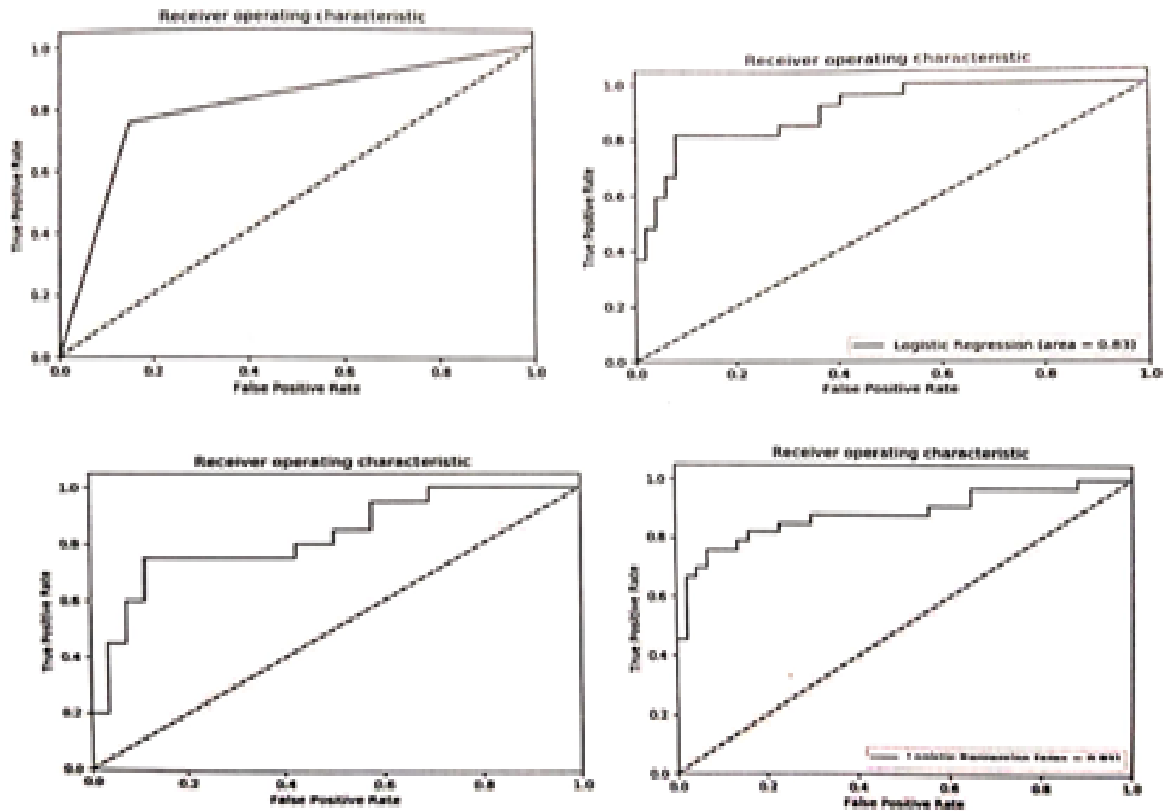


Fig. 13 Different ROC curves based on Decision Tree, Logistic Regression, Naive Bayes and SVM algorithms

5. CONCLUSION AND FUTURE SCOPE

This application helps in faster diagnosis, also reduces medical error and is easy to apply. But this project should only be considered as an assistive tool and must be completely relied upon only after it has been validated in the future by integrating more practical data. It has been concluded that in the making of the predictive model only some success is achieved and hence there is a need for bringing together various techniques to increase the accuracy of prediction. With the increased database if larger dataset is provided to machine, much accurate prediction would prevail. Testing different prediction techniques, different decision tree types namely information gain and gain ratio and multiple classification voting technique can be made use of. Willing to explore several rules such as association rule, logistic regression and clustering algorithms.

6. REFERENCES

- [1] Palaniappan, S. and Awang, R., 2008, March. Intelligent heart disease prediction system using data mining techniques. In 2008 IEEE/ACS international conference on computer systems and applications (pp. 108-115). IEEE.
- [2] Sultana, M., Haider, A. and Uddin, M.S., 2016, September. Analysis of data mining techniques for heart disease prediction. In 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT) (pp. 1-5). IEEE.
- [3] Nikan, S., Gwady-Sridhar, F. and Bauer, M., 2016, December. Machine learning application to predict the risk of coronary artery atherosclerosis. In 2016 International conference on computational science and computational intelligence (CSCI) (pp. 34-39). IEEE.
- [4] Dangare, C.S. and Apte, S.S., 2012. Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), pp.44-48.
- [5] Parthiban, L. and Subramanian, R., 2008. Intelligent heart disease prediction system using CANFIS and genetic algorithm. International Journal of Biological, Biomedical and Medical Sciences, 3(3).
- [6] Pattekari, S.A. and Parveen, A., 2012. Prediction system for heart disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences, 3(3), pp.290-294.
- [7] Subbalakshmi, G., Ramesh, K. and Rao, M.C., 2011. Decision support in heart disease. Indian Journal of Computer Science and Engineering (IJCSE), 2(2), pp.170-176.
- [8] Taneja, A., 2013. Heart disease prediction system using data Oriental Journal of Computer science and technology, 6(4), pp.457-466.
- [9] Chitra, R. and Seenivasagam, V., 2013. Review of heart disease prediction system using data mining and hybrid intelligent techniques. ICTACT journal on soft computing, 3(04), pp.605-09