International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 27 Issue 1 – APRIL 2020.

HEALTHCARE PREDICTOR USING RANDOM FOREST DECISION TREE ALGORITHM WITH USER INTERFACE

Dr. K.Kartheeban[*1], *Vamshikrishna Bandari*[#2], *Malasree Rallapalli*[#3], *Nagalakshmi Pabbisetty*[#4]

[*]*Assistant Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil, India*

[#]*Student ,Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil, India*

[#]*Student ,Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil, India*

[#]*Student ,Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil, India*

Abstract— **Healthcare in India amidst the ongoing COVID-19 epidemic is really crucial and a daunting task ahead of us. Every citizen needs immediate access to proper health guidance for their health condition/situation including maintenance or improvement of health via the prevention, diagnosis, treatment of disease, illness, injury, and other physical and mental impairments in humans. Health care is generally delivered by health professionals (providers or practitioners) in allied health fields. Health care can be done in different stages it may include providing primary care, secondary care, and tertiary care, as well as in public health. Our work on Healthcare Prediction system targets this specific issue by providing health support to the public through an online consultation platform. The system is loaded with data collected from various accredited sources possessing various symptoms, disease or illness. When the user register in the website it allows user to share their symptoms and issues according to that the system processes the data by using appropriate model and guesses the most accurate illness that could be associated with patient's symptoms. On making sure the problem is addressed, direct consultation to a doctor is facilitated with a detailed report if needed by the end user. This area of research is much needed as the ratio of doctors to patients and the affordability to reach and consult a doctor keeps decreasing. Though there are many others who have jumped into this sector/field, they have failed to provide a fool proof system which we are trying to develop by incorporating large sum of reliable data.**

Index Terms—**Supervised Learning, Logistic regression, Support vector machine (SVM), Random forest, Decision tree, GridSearchCV.**

I. INTRODUCTION

For centuries, our world continues to believe that medical services are a basic need of all the people. According to the world bank, Indian government has spent only 1.17 % of GDP towards healthcare. People are evolving more towards solutions which are more reliable, fast and sustainable in terms of cost and resources. The breath-taking pace of change in the way health care is financed and delivered has brought challenges and new activities to all participants in the healthcare system. Healthcare system needs to revamp itself to such solution set. Healthcare services are more important to maintain good health by improving health via the diagnosis, and treatment of disease, injury, and other physical and mental impairments in human beings. Health care is generally delivered by medical professionals (providers or specialists) in health-related fields. Dentistry, midwifery, nursing, medicine, optometry, audiology, pharmacy, psychology, occupational therapy, physical therapy and other health professions are all part of health care. Health care can be done in different stages it may include providing primary care, secondary care, and tertiary care, as well as in public health. Sometimes we have seen situations where someone belonging to us may need doctors help immediately, but they are not available due to some reason. In present existing system there are many problems like frequently we want to visit a doctor even when we come across normal symptoms, illness, injury. It's tedious task for the user to wait for the doctor's appointment long time. generally, people are not aware of type of drugs and medicines that are essential to use for a particular disease. Even people are not much aware about the type of diseases and illness that person get affected. Even for every minor reason also, we have to reach the hospital. The proposing healthcare predictor system is an online consultation project which is an end user support system. It allows users to get instant guidance on their health issues through an intelligent online health care system. The system is fed with various symptoms, illness associated with the user. The intelligent system allows user to share their symptoms and issues. Once the symptoms or issues faced by the user are given, then the system processes user's symptoms to check for different illness, diseases that could be associated with it. We are going to use some intelligent techniques to guess the most accurate illness that could be associated with patient's symptoms.

In these online system ,users can get instant guidance and precautions on their health issues through an intelligent health care system by using SVM(support vector machine algorithm), Random forest algorithm, logistic regression model Unified Modelling Language (UML) is used as a dialect for indicating, visualising, constructing and archiving the curios of programming, non-programming frameworks and for business demonstration. The system is fed with various symptoms and the disease/illness associated with those systems. The system allows user to share their symptoms and issues. Once the symptoms or issues faced by the user are given, then the system processes user's symptoms to check for different illness, diseases that could be associated with it using Machine learning algorithm, it guesses the most accurate illness that could be associated with patient's symptoms and issues. If the healthcare predictor system is not able to provide or predict suitable results, it intimates the user about the type of illness/disease or disorder it feels with the associated user's symptoms /issues. If user's symptoms are unable to match exactly with various disease in our database, the system shows the diseases user might probably get attacked by considering his/her symptoms. These online systems have another benefit it provides the user with doctor address, contact number along with Feedback and administrator dashboard for system operations.

II. LITERATURE SURVEY

Basically, Literature survey is the one of the most important step or stage in software development process. Yanwei Xing, Jie Wang and Zhihong Zhao published the *"Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease"* - From decades the prediction and the survival of Coronary heart disease (CHD) has been a challenging research problem for every medical society. The main goal of this paper is to develop the best data mining algorithms for predicting the survival of CHD patients based on sample 1000 cases. During these processes we carry out a clinical observation and a 6-month follow up to collect the information about the 1000 CHD cases. The survival of disease information of each case is obtained via a proper follow up. Based on this obtained data, we have employed three most popular data mining algorithms to develop, implement the most prediction models using the sample 502 cases. We used the 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes. According to the results we obtained that the SVM is the best predictor with a performance 92.1 % accuracy on the holdout sample of artificial neural networks came out to be the second best with 91.0% accuracy and the third model decision trees models came out to be the worst of the three with 89.6% accuracy. This different comparative study using multiple prediction models for the prediction of CHD patients along with method named as a 10-fold cross-validation which will help us with a proper insight into the relative prediction ability of different data. Gitanjali J. et al., "*apriori algorithm based medical data mining for frequent disease identification"*- The data mining techniques plays a vital role in the process of analysing, extracting a huge amount of data from different sources and then summarizing it into the useful information. This obtained information can be transformed into the knowledge obtained from the historical patterns and future analysis, trends. Here Data mining plays a crucial and significant role in the field of information technology. Health care sector these days generates a huge amount of data about the patient's details, symptoms, issues, hospitals resources, diagnosis methods, electronic patients' records, etc. The data mining techniques are more useful to make the medicinal decisions in the way to cure diseases. The healthcare sector, industry generally collects a huge amount of data which is unfortunately, cannot be "mined" to discover the hidden patterns, information for proper decision making. This discovered knowledge can be used by the healthcare administrators, industry to improve and provide the quality of service. In this paper the authors developed a method, model to identify the frequency of diseases in particular area in a given time period with the aid of association rule based on the Apriori data mining technique.

III. METHODOLOGY

The proposed system allows users to get instant guidance on their health issues through an intelligent online health care system. The system is fed with various symptoms, illness associated with the user. The intelligent system allows user to share their symptoms and issues. Once the system receives the symptoms of user it processes user's symptoms to check different illness that might be associated with patient's symptoms. To predict the accurate result, we use intelligent techniques. The intelligent model predicts the most accurate illness/disease that might be associated with patient's symptoms /issues. Here the model is trained on labelled data to predict whether the disease is present or not. The machine learning model used here is ensemble learning model. In that we considered a random forest or random decision tree. The model is trained with various types of symptoms that are associated with different diseases and used cross fold validation method to validate the model.

IV. IMPLEMENTATION

The entire project is implemented in three modules - User, Admin and Doctor.

***User*:** In this module User can register regarding the basic details like username, password, email, phone, etc. which are mandatory to identify the disease. After that user can login into the website using username and password. If any appropriate data is encountered the system throws a printed message like incorrect password, please fill the rows with correct details etc. Once the user completes the login process the User is allowed to fill the details regarding the symptoms/issues. According to that the System will Predict the particular disease by providing with a proper precaution, instructions to follow for curing the disease. User can have different options to search for example Dengue Fever, Malaria, typhoid, heart attack. According to user symptoms the system will Predict. the disease by providing proper precautions. User have another benefit he/her can search for doctor to talk about their illness/injury and will be provided with instant diagnosis.

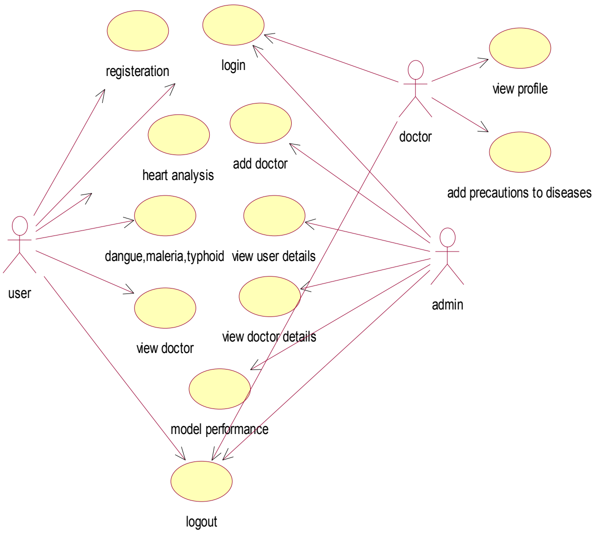


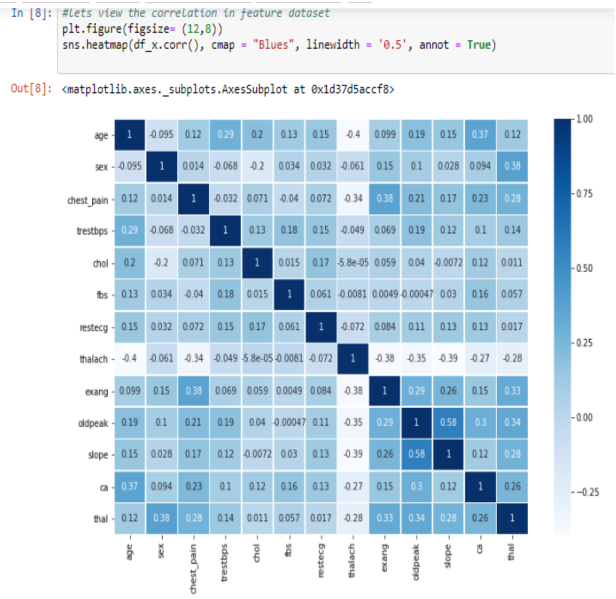
Figure 1 – implementation use case diagram

Admin: In this module admin can login into the website by providing the username and password. Admin can add information about doctor and can view user, doctor details whenever needed. Admin can Upload the Independent and Dependent Train and Test Data sets by Selecting the different algorithm s namely Random forest Algorithm, Support vector machine algorithm. This algorithm when trained on the labelled dataset gives the accuracy results.

Doctor: In this module doctor can login by providing details like username and password. Doctor can view the profile and can suggest the precautions necessary for a disease.

V. MODEL DEPLOYMENT

A. Data Set: The data sets used in this project are collected from Kaggle website. This is the place where we can notice different types of projects, datasets, current ongoing works. There are multiple number of datasets with different information. In this we are involved in exploring the data how it is correlated with other variables in the data sets mainly with independent values. The below graph describes the correlation between attributes.



```
In [8]: #lets view the correlation in feature dataset
        plt.figure(figsize= (12,8))
        sns.heatmap(df_x.corr(), cmap = "Blues", linewidth = '0.5', annot = True)

Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x1d37d5accf8>
```

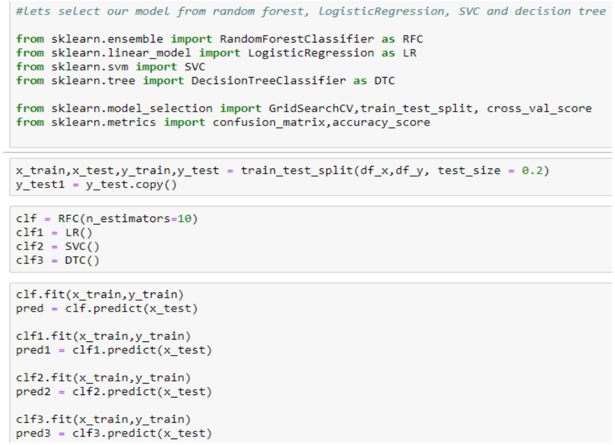
	age	sex	chest_pain	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
age	1	-0.095	0.12	0.29	0.2	0.13	0.15	-0.4	0.099	0.19	0.15	0.37	0.12
sex	-0.095	1	0.014	-0.068	-0.2	0.034	0.032	-0.061	0.15	0.1	0.028	0.094	0.38
chest_pain	0.12	0.014	1	-0.032	0.071	-0.04	0.072	-0.34	0.38	0.21	0.17	0.23	0.28
trestbps	0.29	-0.068	-0.032	1	0.13	0.18	0.15	-0.049	0.069	0.19	0.12	0.1	0.14
chol	0.2	-0.2	0.071	0.13	1	0.015	0.17	-5.8e-05	0.059	0.04	-0.0072	0.12	0.011
fbs	0.13	0.034	-0.04	0.18	0.015	1	0.061	-0.0081	0.0049	-0.00047	0.03	0.16	0.057
restecg	0.15	0.032	0.072	0.15	0.17	0.061	1	-0.072	0.084	0.11	0.13	0.13	0.017
thalach	-0.4	-0.061	-0.34	-0.049	-5.8e-05	-0.0081	-0.072	1	-0.38	-0.35	-0.39	-0.27	-0.28
exang	0.099	0.15	0.38	0.069	0.059	0.0049	0.084	-0.38	1	0.29	0.26	0.15	0.33
oldpeak	0.19	0.1	0.21	0.19	0.04	-0.00047	0.11	-0.35	0.29	1	0.58	0.3	0.34
slope	0.15	0.028	0.17	0.12	-0.0072	0.03	0.13	-0.39	0.26	0.58	1	0.12	0.28
ca	0.37	0.094	0.23	0.1	0.12	0.16	0.13	-0.27	0.15	0.3	0.12	1	0.26
thal	0.12	0.38	0.28	0.14	0.011	0.057	0.017	-0.28	0.33	0.34	0.28	0.26	1

Figure 2 - correlation graph of feature vector

The above figure describes about correlation if the blue colour has more intensity then it has more correlated between x and y attributes. The diagonals are a pair of similar attributes so they are more correlated in the graph compared to others.

b. Model selection: The dependent variable used here is categorical. So, we opt for supervised classification models. When it comes to classification, we have multiple machine learning models to deploy in Scikit-learn.

We used logistic regression, support vector machine, decision tree and ensemble learning method namely random decision tree.



```
#Lets select our model from random forest, LogisticRegression, SVC and decision tree

from sklearn.ensemble import RandomForestClassifier as RFC
from sklearn.linear_model import LogisticRegression as LR
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier as DTC

from sklearn.model_selection import GridSearchCV,train_test_split, cross_val_score
from sklearn.metrics import confusion_matrix,accuracy_score
```

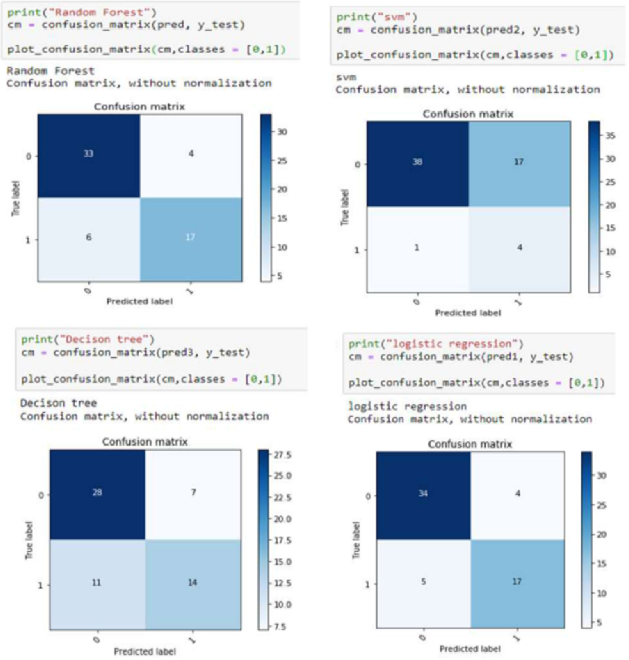
```
x_train,x_test,y_train,y_test = train_test_split(df_x,df_y, test_size = 0.2)
y_test1 = y_test.copy()
```

```
clf = RFC(n_estimators=10)
clf1 = LR()
clf2 = SVC()
clf3 = DTC()
```

```
clf.fit(x_train,y_train)
pred = clf.predict(x_test)

clf1.fit(x_train,y_train)
pred1 = clf1.predict(x_test)

clf2.fit(x_train,y_train)
pred2 = clf2.predict(x_test)

clf3.fit(x_train,y_train)
pred3 = clf3.predict(x_test)
```

Figure 3 - deploying Classification models

Despite of all these four models we choose to use random forest which gives more accuracy on testing data set. The accuracy and the confusion matrixes are shown below in pictorial way. The graph shows the confusion matrixes of different models obtained with the help of matplotlib library.



```
print("Random Forest")
cm = confusion_matrix(pred, y_test)

plot_confusion_matrix(cm,classes = [0,1])
```

```
Random Forest
Confusion matrix, without normalization
```

```
print("svm")
cm = confusion_matrix(pred2, y_test)

plot_confusion_matrix(cm,classes = [0,1])
```

```
svm
Confusion matrix, without normalization
```

```
print("Decison tree")
cm = confusion_matrix(pred3, y_test)

plot_confusion_matrix(cm,classes = [0,1])
```

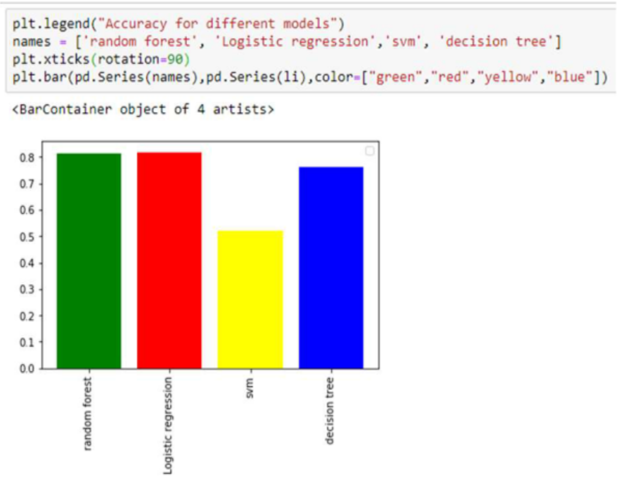
```
Decison tree
Confusion matrix, without normalization
```

```
print("logistic regression")
cm = confusion_matrix(pred1, y_test)

plot_confusion_matrix(cm,classes = [0,1])
```

```
logistic regression
Confusion matrix, without normalization
```

Figure 4 – confusion matrixes for 4 different

From the above graph the confusion matrix works fine with the random forest and logistic regression which is almost similar to the dataset. The accuracy is plotted in bar graph and also showed in tabular format before we start tuning the hyperparameters on the same dataset.



```
plt.legend("Accuracy for different models")
names = ['random forest', 'Logistic regression','svm', 'decision tree']
plt.xticks(rotation=90)
plt.bar(pd.Series(names),pd.Series(li),color=["green","red","yellow","blue"])
```

```
<BarContainer object of 4 artists>
```

Figure 5 - Accuracy scores of 4 models

Hence it is clear that random forest which is indicated in green colour and logistic regression which is indicated in red colour are almost works fine for data set with similar accuracy.

Table 1 - Accuracy scores

S No.	***Model Name***	***Accuracy score***
1	***Random Forest***	***82.12%***
2	***Logistic Regression***	***82.12%***
3	***Support Vector Machine***	***54.03%***
4	***Decision Tree***	***76.58%***

c. Random forest Model: In spite of using all supervised classification models we preferred to use the ensemble learning method which is best in predicting accuracy on training and testing datasets. So, we implement ensemble learning method especially for random forest decision tree to make our model more intelligent and powerful. The model starts building multiple trees on sample data set and whenever a new data record comes to the model it performs voting on multiple decision trees that gives the result based upon on voting which have the gained the best votes it gives that as a result. When the model is trained it takes multiple features as consideration and builds the multiple different decision trees by varying its features and performs voting on new predicted data record and gives result with maximum votes.

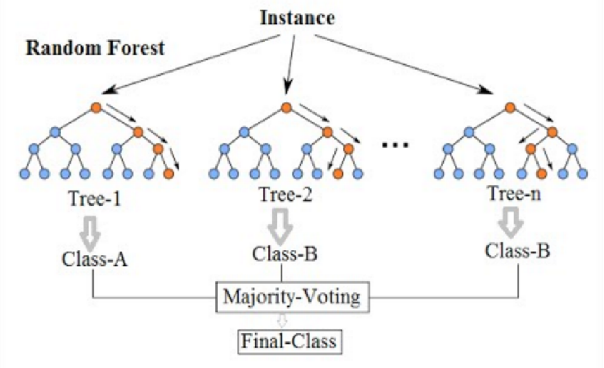


Figure 6 - random forest

d. Properties of Random Forest:

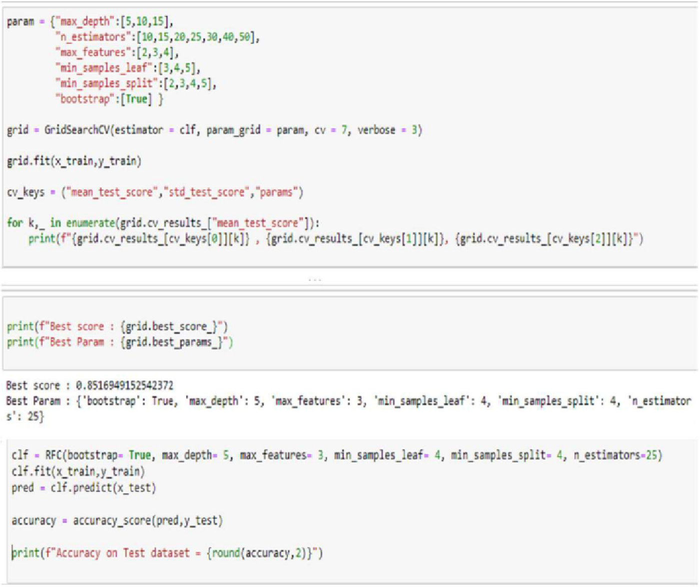
- Random forest is a predictive modelling algorithm.
- Here we used random forest for classification.
- It works well with hyper-parameters.
- The correlation between any two trees in the forest.
  - Increasing the correlation will automatically increases the forest error rate.
- Generally, we consider a tree with a low error rate to be a strong classifier.
  - Increasing the strength of the individual trees decreases the forest error rate It runs efficiently on large and small datasets.

e. Algorithm:

- Import all necessary libraries.
- Import data set using Pandas data frame.
- Perform data mining on data frame.
- Manipulate the data and remove outliers from the data.
- Train the model in a way that:
  - Randomly select ‘m’ features from the total features.
  - Select the root node using best split and form various decision trees.

- Predict the outcome using these decision trees.
- Calculate the vote for each of the target predicted by each tree.
- The target with the highest vote is considered as the final prediction of the random forest algorithm.

f. Tuning Hyperparameters: When we consider the random forest as our model and tunes its hyper parameters by using GridSearchCV - a algorithm in scikit-learn library that used to perform all models on data set by varying its hyperparameters that provides and gives the best accuracy model with its hyperparameters.



```
param = {"max_depth":[5,10,15],
         "n_estimators":[10,15,20,25,30,40,50],
         "max_features":[2,3,4],
         "min_samples_leaf":[3,4,5],
         "min_samples_split":[2,3,4,5],
         "bootstrap":[True] }

grid = GridSearchCV(estimator = clf, param_grid = param, cv = 7, verbose = 3)

grid.fit(x_train,y_train)

cv_keys = ("mean_test_score","std_test_score","params")

for k,_ in enumerate(grid.cv_results_["mean_test_score"]):
    print(f"{grid.cv_results_[cv_keys[0]][k]} , {grid.cv_results_[cv_keys[1]][k]}, {grid.cv_results_[cv_keys[2]][k]}")
```

```
print(f"Best score : {grid.best_score_}")
print(f"Best Param : {grid.best_params_}")
```

```
Best score : 0.8516949152542372
Best Param : {'bootstrap': True, 'max_depth': 5, 'max_features': 3, 'min_samples_leaf': 4, 'min_samples_split': 4, 'n_estimator
s': 25}
```

```
clf = RFC(bootstrap= True, max_depth= 5, max_features= 3, min_samples_leaf= 4, min_samples_split= 4, n_estimators=25)
clf.fit(x_train,y_train)
pred = clf.predict(x_test)

accuracy = accuracy_score(pred,y_test)

print(f"Accuracy on Test dataset = {round(accuracy,2)}")
```

Figure 7 - random forest hyperparameter tuning

The best hyperparameters is provided by GridSearchCV are bootstrap is True, maximum depth of nodes are five, the maximum features to take are 3, minimum samples at leaf are 4, split on samples test is 4 and number of estimators to use are 25. By providing these it gives an accuracy on validation and training set is 85.169% and on test data is 83.92% accuracy.

VI. USER INTERFACE

The entire project is implemented in three modules with User friendly interfaces using web technologies. The interfaces are User, Admin and Doctor.

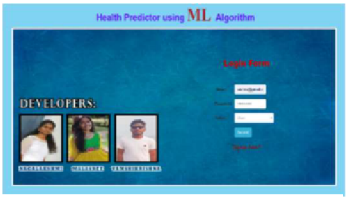


Figure 8 - UI Login

For every interface there must be two things in common to provide security to the users. The sign in page. When user enter the system after signup, he/her needs to login to enter the working modules. Where user can actually see two modules in prediction. One is for heart analysis and another is for normal health issues like dengue, typhoid, malaria.

The two diagrams here describe two random forest models prediction using two datasets. The user needs to fill the above form for reach their respective field. Once they fill all fields which are required then they can submit the form by clicking on predict .it means the model will predict as per data provided by user and it provides the precautions for the predicted disease as shown below. After prediction of even the user can book the appointment by clicking on the button which is named as book appointment. To get a proper medication, precautions user can register to the doctors which is available in our site.

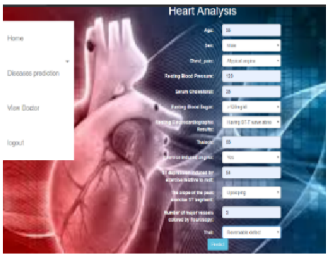


Figure 9 - heart analysis prediction

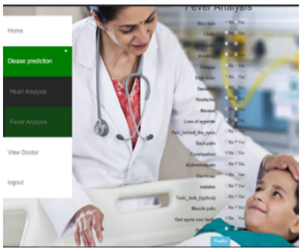
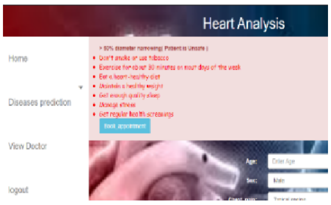


Figure 10 - fever predictions



Heart Analysis

> 50% diameter narrowing(Patient is Unsafe)

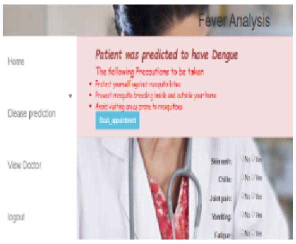
- Don't smoke or use tobacco
- Exercise for about 30 minutes on most days of the week
- Eat a heart-healthy diet
- Maintain a healthy weight
- Get enough quality sleep
- Manage stress
- Get regular health screenings

Book appointment

Age: Enter Age
Sex: Male

Figure 11 - heart precautions

After clicking the button, a new page opens for booking the appointment. user can cross check in the dashboard for different doctors in our website. User has an opportunity to make telephonic call to a particular doctor which are available in the site as shown below. Once user is done with his/her work they can logout from the interface by clicking on logout menu icon.



Fever Analysis

Patient was predicted to have Dengue

The following Precautions to be taken

- Protect yourself against mosquito bites
- Prevent mosquito breeding inside and outside your home
- Avoid visiting areas prone to mosquitoes

Book_appointment

Skin rash: ○No ○Yes
Chills: ○No ○Yes
Joint pain: ○No ○Yes
Vomiting: ○No ○Yes
Fatigue: ○No ○Yes

Figure 12 - fever precautions

VII. CONCLUSION

The system would drastically reduce the human effort, and it reduces the cost, time constraint in terms of human resources and expertise by increasing the diagnostic accuracy. Generally, the prediction of disease using Data Mining applications is a challenging and risky task as the data found to be noisy, irrelevant and massive too. In this scenario, data mining tools come in handy in exploring the knowledge of the medical data which is quite interesting.

Advantages:

- Healthcare Prediction system helps in reducing the cost of medical tests as users has to spend huge amount of money on diagnostics centres.
- User can search for doctor's help at any point of time and instantly.
- They can clarify their doubts regarding the symptoms that prevails in their body in an instance.
- Disease diagnosis time is less.
- User can talk about their illness openly to machine model.
- Doctors get more clients through online.
- People will become familiar with medicines and drugs that they are consuming.
- User will be clear about the precautions for particular type of diseases.
- Users get more knowledge when they use these intelligent models.
- Users can obtain doctor's appointment easily.

Future Scope: The current work generally helps to open up a new research area. The models that we used to predict the disease based on user symptoms can be further extended by considering different types of issues/diseases. There are many possible improvements that can be explored to improve the scalability and accuracy of the prediction system. Currently we have developed a generalized system. In future we can use these intelligent systems with different data sets containing different symptoms and diseases to train the model. The performance and the accuracy of the health care can be improved significantly by handling numerous class labels, data sets in the prediction process. This will lead to a positive impact in the field of medical research. So, when the model is trained with different datasets, we can extract powerful insights because the datasets are extremely enormous. Generally, the dimensionality of the heart disease database is high and thus the identification, selection of significant attributes for better accuracy and diagnosis of heart disease are the challenges for future research.

REFERENCES

[1] Shaikh Abdul Hannan, A.V. Mane, R. R. Manza, and R. J. Ramteke, Dec 2010, "Prediction of Heart Disease Medical Prescription using Radial Basis Function", IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), DOI: 10.1109/ICCIC.2010.5705900 ,28-29 .

[2] Mrudula Gudadhe, Kapil Wankhade, and Snehlata Dongre, Sept 2010,"Decision Support System for Heart Disease Based on Support Vector Machine and Artificial Neural Network", International Conference on Computer and Communication Technology (ICCCT),DOI:10.1109/ICCCT.2010.5640377, 17-19.

[3] http://www.heart.org/HEARTORG/Conditions/HeartAttack/WarningSignsofa HeartAttack/Warning-Signs-of-aHeartAttack_UCM_002039_Article.jsp#.WNpKgPl97IU.

[4] www.who.int/cardiovascular_diseases/en/. http://food.ndtv.com/health/world-heart-day-2015-heart-disease-in-india-is-agrowing-concern-ansari-1224160.

[5] https://en.wikipedia.org/wiki/Cardiovascular_disease.

[6] AH Chen, SY Huang, PS Hong, CH Cheng, and EJ Lin,2011, "HDPS: Heart Disease Prediction System", Computing in Cardiology, ISSN: 0276-6574, pp.557-560.

[7] Manpreet Singh, Levi Monteiro Martins, Patrick Joanis, and Vijay K. Mago,2016, "Building a Cardiovascular Disease Predictive Model using Structural Equation Model & Fuzzy Cognitive Map", IEEE International Conference on Fuzzy Systems (FUZZ), pp. 1377-1382.

[8] Kaggle – A data set collection hub.

[9] "Health Topics: Health Systems" www.who.int. WHO World Health Organization. Retrieved 2013-11-24.

[10] "Health at a Glance 2013 - OECD Indicators" (PDF). OECD. 2013-11-21. pp. 5, 39, 46, 48. Retrieved 2013-11-24.

[11] "OECD. StatExtracts, Health, Health Status, Life expectancy, Total population at birth, 2011" (online statistics). stats.oecd.org/. OECD's iLibrary. 2013. Retrieved 2013-11-24.

[12] World Health Organization. Anniversary of smallpox eradication. Geneva, 18 June 2010.

[13] United States Department of Labor. Employment and Training Administration: Health care. Retrieved June 24, 2011.