Disease Prediction System using Naïve Bayes algorithm

Department Of Computer Science & Application, College Of Engineering And Technology
(Techno Campus ,Po-Ghatikia, Mahalaxmi Vihar, Bhubaneswar - 751029, India)
cet.edu.in

Abstract. In this era of IT, technology has revolutionized the health domain to a great extent. This project aims to design a diagnostic model for various diseases relying on their symptoms. This System has used data mining techniques such as classification in order to achieve such a model. Datasets consisting of voluminous data about patient diseases are gathered, refined and classified and were used for training the intelligent agent. Here, the Naive Bayes Algorithm is used for classification purpose. Naïve Bayes Classifier calculates the probability of the disease. Based on the result, the patient can contact the doctor accordingly for further treatment. It's an exemplar where technology and health knowledge are sewn into a thread perfectly with a desire to achieve "prediction is better than cure".

Keywords: Naive bayes, medical data, classification, data mining

1 Introduction

Nowadays, the use of the internet has been stimulating curiosity among people and, be it of any kind, they are trying to find a solution to their problems through the internet only. It is a matter of fact that people have much easier access to the internet than hospitals and doctors. So, with the help of this system, a user can consult a doctor by sitting at their home itself. There will not be any fuss of visiting a clinic or hospitals and making your health condition worse. This Disease Prediction system is a web-based application that predicts the most probable disease of the user in accordance with the given symptoms by the help of the data-sets collected from different health-related sites. It often happens that someone nearer or dearer to you may need a doctor's help immediately for some serious reasons but the doctor isn't available for consaltation for some prior commitments or other obvious reasons. That is when the role of this automated program comes into play. This Disease Prediction system can be used for urgent guidance on their illness according to the details and symptoms they will feed to the web-based application. Here, we use some intelligent data processing techniques to get the most accurate disease that would be related to the patient's details. And then based on the results, the patient can contact the respective disease specialist for any further treatments. This system can be used for a free consultation regarding any illness.

2. Literature survey

Research for the best and simplest medical diagnosis mining technique had been carried out by Al-Aidaroos K.M. et. al [1]. The authors have compared For the Naïve Bayes with five other classifiers, i.e. Logistic Regression (LR), KStar (K*), Decision Tree (DT), Neural Network (NN) and a simple rule-based algorithm (ZeroR). They have used 15 real-world medical problems from the UCI machine-learning repository [2] those were selected for evaluation. It was found that Naïve Bayes outperforms the other algorithms in 8 out of 15 data sets and hence was considered as a better predictive technique than others as shown in fig. 1.

Jyoti Soni et. al. have conducted number of experiments to compare the performance of predictive data mining technique on the same dataset and the outcome reveals that Decision Tree outperforms and sometime Bayesian classification is having similar accuracy as of decision tree but other predictive methods like KNN, Neural Networks, Classification based on clustering are not performing well [3]. Also, they suggested that the accuracy of the Decision Tree and Bayesian Classification further improves after applying genetic algorithm to reduce the actual data size to get the optimal subset of attribute sufficient for heart disease prediction.

Table 1- Predictive Accuracy of Bayes and other Technique

Medical Problems	NB	LR	K*	DT	NN	ZeroR
Breast Cancer wise	97.3	92.98	95.72	94.57	95.57	65.52
Breast Cancer	72.7	67.77	73.73	74.28	66.95	70.3
Dermatology	97.43	96.89	94.51	94.1	96.45	30.6
Echoeardiogram	95.77	94.59	89.38	96.41	93.64	67.86
Liver Disorders	54.89	68.72	66.82	65.84	68.73	57.98
Pima Diabetes	75.75	77.47	70.19	74.49	74.75	65.11
Haeberman	75.36	74.41	73.73	72.16	70.32	73.53
Heart-c	83.34	83.7	75.18	77.13	80.99	54.45
Heart-statlog	84.85	84.04	73.89	75.59	81.78	55.56
Heart-b	83.95	84.23	77.83	80.22	80.07	63.95
Hepatitis	83.81	83.89	80.17	79.22	80.78	79.38
Lung Cancer	53.25	47.25	41.67	40.83	44.08	40
Lymphpgraphy	84.97	78.45	83.18	78.21	81.81	54.76
Postooerative Patient	68.11	61.11	61.67	69.78	58.54	71.11
Primary tumor	49.71	41.62	38.02	41.39	40.38	24.78
Wins	8\15	5\15	0\15	2\15	1\15	1\15

(Al-Aidaroos, Bakar, & Othman, 2012)

Fig.1 Literature survey

Pattekari S.A. & Parveen A. conducted research to predict heart diseases where the user provides the data, which is compared with trained set of values, and thereafter makes prediction about the heart disease [4]. Masethe H.D. et. al. used data mining algorithms such as J48, Naïve Bayes, REPTREE, CART, and Bayes Net for predicting heart attacks [5]. Their result showed a prediction accuracy of 99% justifying data mining techniques to be used in health sector in order to predict patterns from the dataset.

There are many papers that diagnose heart diseases. Earlier decision tree based systems were used for diagnosing heart diseases. Shouman Mai et. al. evaluated the performance of an alternative Decision Trees a model [6] that outperforms J4.8 Decision Tree and Bagging algorithm in the diagnosis of heart disease patients. They estimated the sensitivity, specificity, and accuracy and found out it to be better than other algorithms. The same researchers found 83.3% accuracy in adding k-means clustering in diagnosis [7]. Vembandasamy, K. et. al. used navie bayes algorithm to classify the medical data and found out 86.4198% of accuracy with minimum time [8].

3. Proposed system

In the proposed system the diseases are predicted automatically by the system using our model which is trained on the medical dataset. This system also shows the confidence score of the prediction. After getting the anticipated disease, the system will suggest doctors associated with that disease and therefore the patient can consult to the doctor online. The proposed system acts as a decision support system and will prove to be an aid for the physicians with the diagnosis.

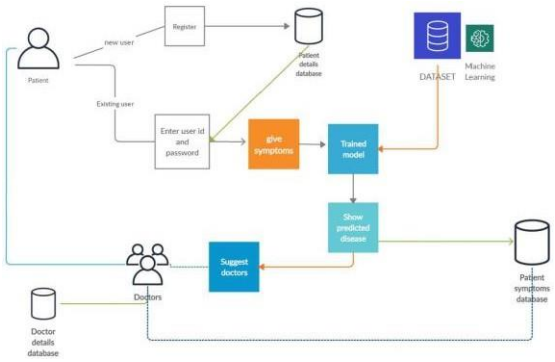

Fig. 2. Architecture of the proposed System

The disease prediction system has 3 users such as doctor, patient and admin. Each user of the system is authenticated by the system. The system allows the patient to give symptoms and according to those symptoms the system will predict a disease with a accuracy score. Then it suggests doctors for online consultation. Then the patient can consult a doctor anytime at his convenience.

4. Algorithm implementation and evaluation:

4.1 Algorithm used

Naive Bayes - There is a wide range of major algorithms for predicting various diseases with guessable symptoms in the field of Supervised Machine Learning and its different models such as Decision tree, Naive Bayes and Random Forest. Naïve Bayes has three different models i.e. Gaussian, Multinomial and Bernoulli Naïve

Bayes. Each model has its own accuracy to predict the result of diseases and the application of data fitting is mostly the same in all the three models. As Gaussian Naive Bayes is comparatively easy to understand and quite simpler than the other two, this project work has been done using the same.

```
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
```

Naïve Bayes classifier depends on Bayes Theorem.
Bayes theorem:

$$P(Y/X1,\ldots\ldots\ldots\ldots,Xn) = \frac{P(Y)\,P(X1,\ \ldots\ldots\ldots,Xn/Y)}{P(X1,\ Xn)}$$

Where,

Y is the class variable

X1 , X2, ... , Xn are the dependent features

4.2. Data collection

Data collection has been done from the internet to identify the disease here the real symptoms of the disease are collected i.e. no dummy values are entered. The dataset is collected from kagggle.com.[9]

This csv file contains 5000 rows and 133 columns, 132 columns for the unique symptoms. And last column for the disease class (40 unique disease classes).
Some rows of disease with their corresponding symptoms in the dataset –

	Disease	Symptoms
0	Malaria	[chills, vomiting, high_fever, sweating, heada...
1	Allergy	[continuous_sneezing, shivering, chills, water...
2	Fungal infection	[skin_rash, nodal_skin_eruptions, dischromic _...
3	Gastroenteritis	[vomiting, sunken_eyes, dehydration, diarrhoea]
4	arthritis	[muscle_weakness, stiff_neck, swelling_joints,...
5	Typhoid	[chills, vomiting, fatigue, high_fever, headac...
6	Hypertension	[muscle_weakness, stiff_neck, swelling_joints,...

Fig. 3. Dataset

4.3 Performance analysis

Confusion matrix-
A confusion matrix is basically a table that is used to describe the performance of a classification model on a group of test data that truth values are known. From the

confusion matrix table, it is clearly seen that the Naïve Bayes algorithm is predicting all the diseases correctly within the test set.
This confusion matrix showing the "actual class" vs 'predicted class" ratios:

```
[[18,  0,  0, ...,  0,  0,  0],
 [ 0, 30,  0, ...,  0,  0,  0],
 [ 0,  0, 24, ...,  0,  0,  0],
 ...,
 [ 0,  0,  0, ...,  6,  0,  0],
 [ 0,  0,  0, ...,  0, 22,  0],
 [ 0,  0,  0, ...,  0,  0, 34]]
```

Classification report-
Classification report visualizes the precision, recall and f1 score of a model.

Classification report :

	precision	recall	f1-score	support
(vertigo) Paroymsal Positional Vertigo	1.00	1.00	1.00	37
AIDS	1.00	1.00	1.00	42
Acne	1.00	1.00	1.00	42
Alcoholic hepatitis	1.00	1.00	1.00	40
Allergy	1.00	1.00	1.00	36
Arthritis	1.00	1.00	1.00	42
Bronchial Asthma	1.00	1.00	1.00	48
Cervical spondylosis	1.00	1.00	1.00	37
Chicken pox	1.00	1.00	1.00	38
Chronic cholestasis	1.00	1.00	1.00	31
Common Cold	1.00	1.00	1.00	34
Dengue	1.00	1.00	1.00	46
Diabetes	1.00	1.00	1.00	35
Dimorphic hemmorhoids(piles)	1.00	1.00	1.00	50
Drug Reaction	1.00	1.00	1.00	38
Fungal infection	1.00	1.00	1.00	33
GERD	1.00	1.00	1.00	43
Gastroenteritis	1.00	1.00	1.00	43
Heart attack	1.00	1.00	1.00	42
Hepatitis B	1.00	1.00	1.00	47
Hepatitis C	1.00	1.00	1.00	40
Hepatitis D	1.00	1.00	1.00	38
Hepatitis E	1.00	1.00	1.00	50
Hypertension	1.00	1.00	1.00	37
Hyperthyroidism	1.00	1.00	1.00	42
Hypoglycemia	1.00	1.00	1.00	44
Hypothyroidism	1.00	1.00	1.00	38
Impetigo	1.00	1.00	1.00	36
Jaundice	1.00	1.00	1.00	37
Malaria	1.00	1.00	1.00	35
Migraine	1.00	1.00	1.00	39
Osteoarthristis	1.00	1.00	1.00	30
Paralysis (brain hemorrhage)	1.00	1.00	1.00	38
Peptic ulcer diseae	1.00	1.00	1.00	31
Pneumonia	1.00	1.00	1.00	46
Psoriasis	1.00	1.00	1.00	33
Tuberculosis	1.00	1.00	1.00	40
Typhoid	1.00	1.00	1.00	41
Urinary tract infection	1.00	1.00	1.00	41
Varicose veins	1.00	1.00	1.00	40
hepatitis A	1.00	1.00	1.00	44
accuracy			1.00	1624
macro avg	1.00	1.00	1.00	1624
weighted avg	1.00	1.00	1.00	1624

Fig. 4. Classification report

Accuracy score: Our result showed a prediction accuracy of 1.0.

5. Result

The below figure no 5 shows the homepage of our system-

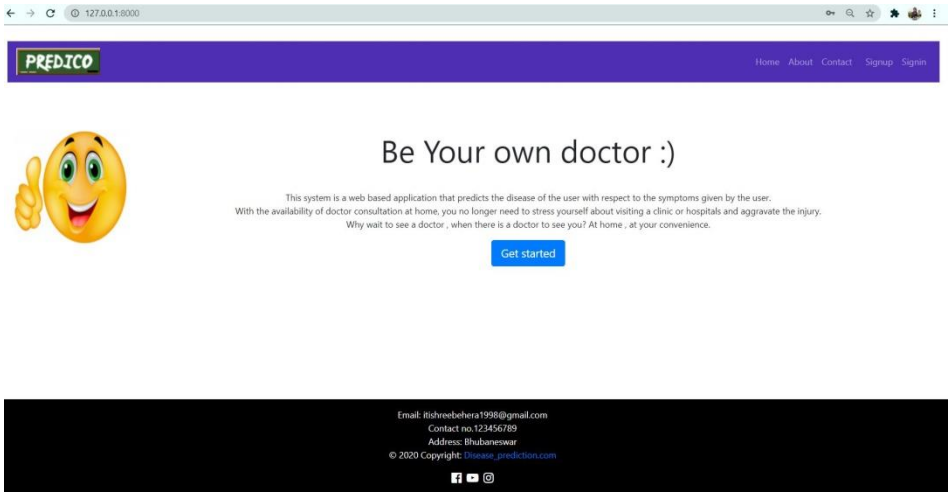

Fig. 5. Home page

Patients and doctors are allowed to register them in the system; after registering they can login to the system by entering valid credentials.
They have to choose their role of accessing the system, either doctor or patient.
The below fig no 6 shows the login page:

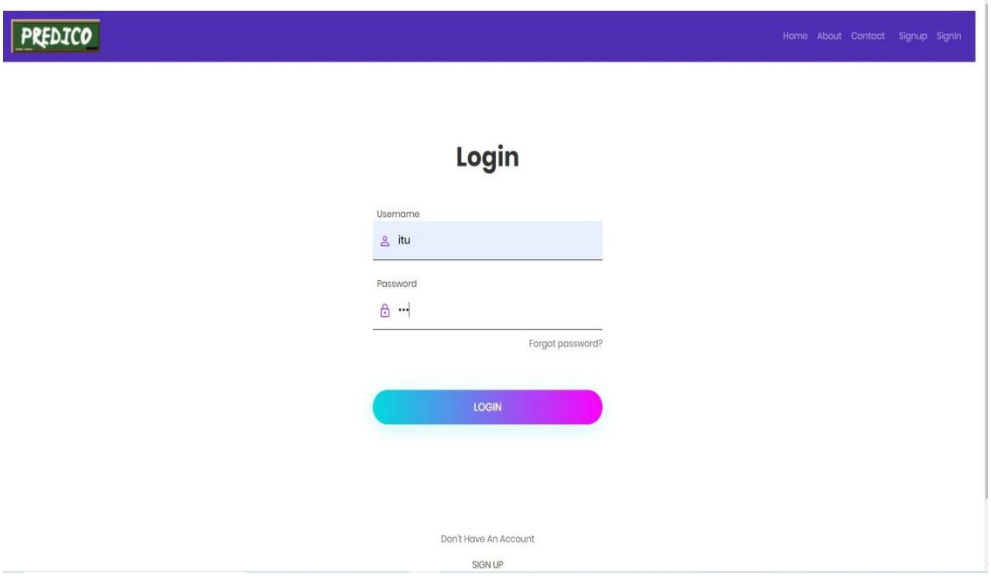

Fig. 6 Login page

After logging in, the users will be directed to their profile page: this is showing patient's profile page-
The patient profile page is having some functionality like:
Check the predicted disease by providing symptoms.
Give feedback to the system.
Check the consultation history.

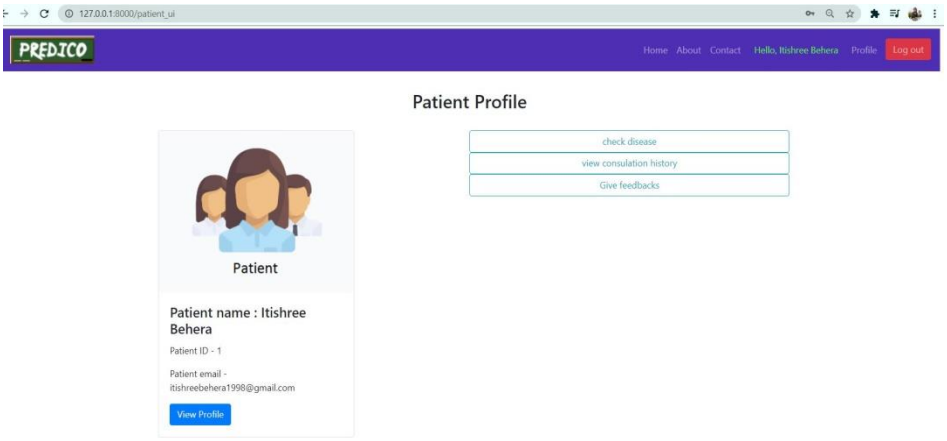

Fig. 7. Patient's profile page

The patients can choose the symptom to get the predicted disease.

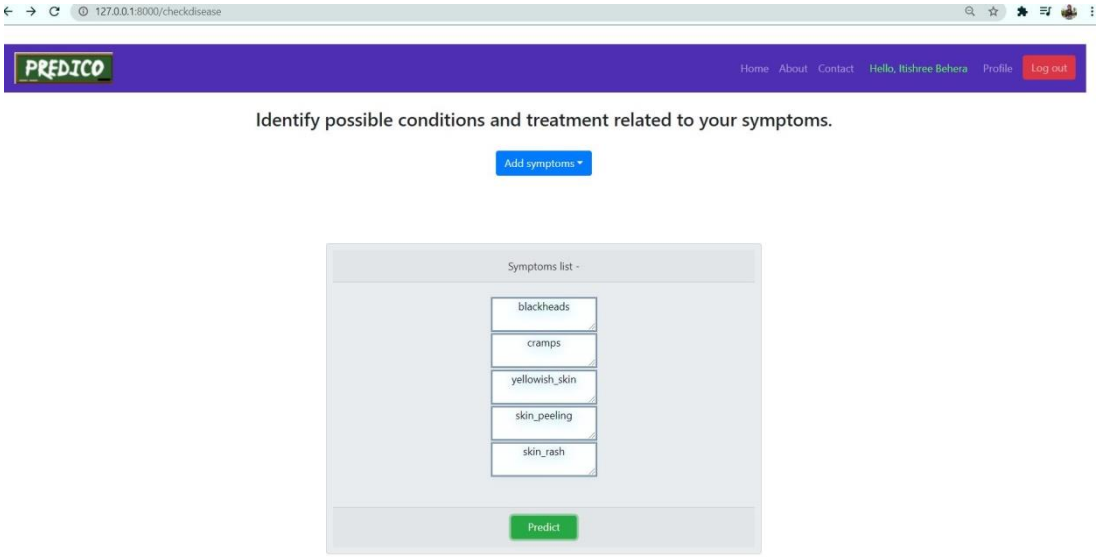

Fig.8 Entering symptoms using a symptom search interface:

The system then gives the predicted disease and suggests the respective specialist:

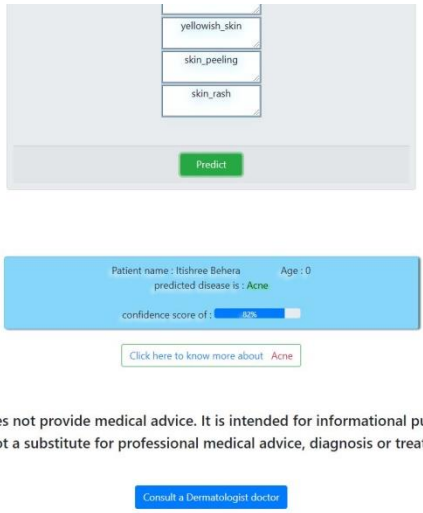

Fig. 9 System gives a predicted disease with an associated doctor specialist to consult

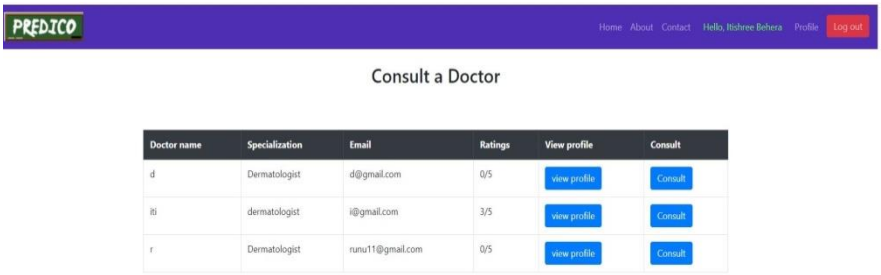

Fig. 10 Consult a doctor

The below figure no.11 shows the ongoing consultation between a patient and a doctor.

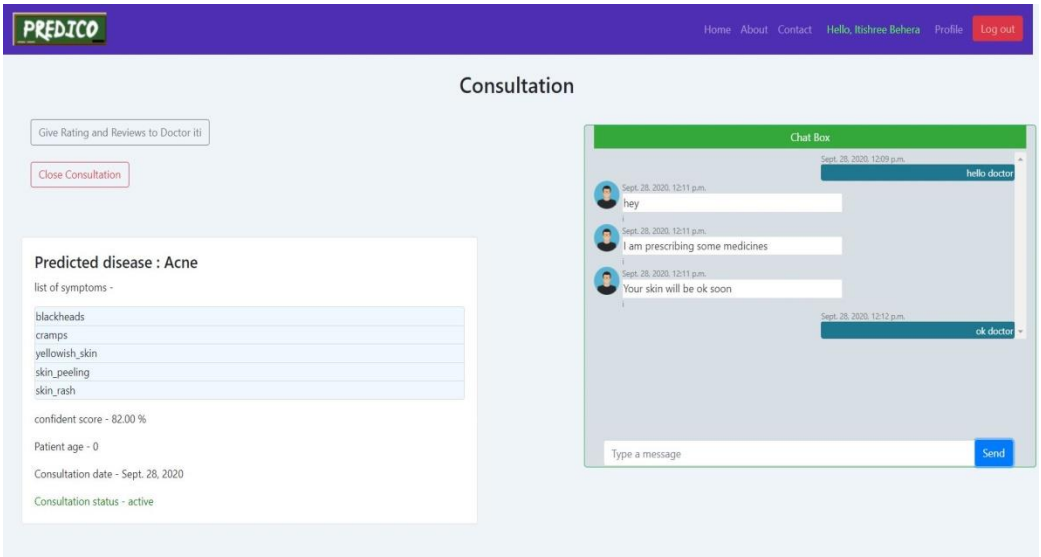

Fig. 11 Consultation UI

The user can see the consultation history anytime and give the feedback to the doctor and the system as well .That will help the admin improving the functionalities of the system.

7. Conclusion

Our proposed methodology, Disease prediction system intended to give better output results. We found about 100% accuracy on our dataset which is a more than any existing systems. This system will provide an up to the mark assistance regarding your current health status on the basis of a small survey of personal details. This project aims to predict the disease on the basis of the symptoms. The project is designed in such a way that the system takes symptoms from the user as input and predicts disease as an output. Because of our system, Patients won't have to wait for Doctors appointments, due to our system patients save their money and time. After getting the anticipated disease, the system will suggest doctors associated with that disease and therefore the patient can consult the doctor online. The proposed system acts as a decision support system and will prove to be an aid for the physicians with the diagnosis.

8. Future work

It is impossible to develop a system that makes all the requirements of the user. As the system is being used, the user's requirement changes. In the future, we'll add more Symptoms and our system will predict for more disease. Thus in this paper, we've successfully studied the existing architecture of the system and through the proposed architecture we are designing a system that will help in predicting the disease using the user-friendly web application. The system should be efficient to predict the diseases and in giving correct suggestion of doctors to consult using machine learning. Further, the system can be extended to have more number of symptoms and diseases. Currently, it does not recommend medications of the disease and Past history of the disease has not been considered, which can be implemented in the future.

Acknowledgement

We would wish to express our sincere gratitude to our advisor, Dr. Jibitesh Mishra, Associate Professor, Computer Science & Engineering Department, whose knowledge and guidance have motivated us to achieve goals we never thought possible.

We would also like to convey our deep regards to all other faculty members of the Dept. of CSA,CET,BBSR, who have bestowed their great effort and guidance at appropriate times without which it would have been very difficult on our part to finish this work.

References

1. Al-Aidaroos, K.M., Bakar, A.A. and Othman, Z.: Medical data classification with Naive Bayes approach. Information Technology Journal. 11(9), 1166 (2012).
2. Asuncion, A. and Newman, D.: UCI machine learning repository downloaded from https://ergodicity.net/2013/07/.
3. J Soni, J., Ansari, U., Sharma, D. and Soni, S.: Predictive data mining for medical diagnosis: An overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-48 (2011).
4. Pattekari, S.A. and Parveen, A.: Prediction system for heart disease using Naïve Bayes. International Journal of Advanced Computer and Mathematical Sciences. 3(3), 290-294 (2012).
5. Masethe, H.D. and Masethe, M.A.: Prediction of heart disease using classification algorithms. In Proceedings of the world Congress on Engineering and computer Science. 2, 22-24 International Association of Engineers, Francisco (2014).
6. Shouman, M., Turner, T. and Stocker, R.: Using decision tree for diagnosing heart disease patients. In: Proceedings of the Ninth Australasian Data Mining Conference, Volume. 121, 23-30. Association of Computing Machinery, Victoria (2011).
7. Shouman, M., Turner, T. and Stocker, R., 2012. Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients. In: Proceedings of the International Conference on Data Science (ICDATA), pp. 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), Monte Carlo (2012).
8. Vembandasamy, K., Sasipriya, R. and Deepa, E., 2015. Heart diseases detection using Naive Bayes algorithm. International Journal of Innovative Science, Engineering & Technology, 2(9), pp.441-444.
9. https://www.kaggle.com/neelima98/disease-prediction-usingmachine-learning