DECISION SUPPORT SYSTEM FOR MEDICAL DIAGNOSIS USING DATA MINING

[1]Sumitra Sadhukhan, [2]Hitesh Jadhav, [3]Almas Mohammad, [4]Ayesha Sankhe, [5]Kirti Shinde

[1]Assistant Professor, [2]Research scholar

[1,2,3,4,5]Department of Computer Engineering, Rajiv Gandhi Institute of Technology,
Juhu-Versova Link Road, Mumbai-400053

Email: sumitra.sadhukhan@mctrgit.ac.in, hiteshjadhav868@gmail.com, almasfqureshi1@gmail.com, ayeshasankhe@gmail.com, kirtishinde707@gmail.com

***Abstract:* Our research paper focuses on Medical diagnosis by using decision support system for diagnosis of heart disease and diabetes to help doctors. The use of C4.5 algorithm and Naïve-Bayes classifier is proposed to classify disease and for its prediction. It checks and predicts if patients could in the near future encounter risks of having heart disease or diabetes.**

***Keywords:* Medical diagnosis, Heart disease and diabetes, Data Mining.**

1. INTRODUCTION

The World Health Statistics 2018 state Cardiovascular diseases take lives of 17.9 million people every year causing 31% of global deaths. It enlightens on the fact that one in three adults in the world has a raised blood pressure- that cause stroke and heart diseases.

Also that diabetes causes 1.6 million deaths globally which means an estimate of 8.5% in 2015.

Decision support system is used to find relationships between diseases around the population, patient history, symptoms, pathology, family history and test results. Decision support systems takes into account all the parameters to reach an effective and optimal result. Inadequate patient information is considered as a huge setback for diagnosis of patients. C4.5 algorithm and Naïve-Bayes classifier is used for classification of disease and comparison of their effectiveness, correction rate among them and helping with diagnosis process and rate.

2. LITERATURE SURVEY

According to technical paper review, rule based expert systems are the most commonly known type of knowledge-based systems. The knowledge is represented in the form of IF-THEN rules. Expert systems have been developed and applied to many fields. Diagnosis system is a system which can diagnose diseases through checking out the symptoms. A knowledge based online diagnosis system is developed for diagnosis of diseases based on knowledge given by doctors in the system

Clinical decisions are often made based on doctor’s understanding and skill and encounters rather than on the data hidden in the database.

3. PROPOSED SYSTEM

For improving efficiency, we have proposed the use of C4.5 and Naïve Bayes classifier.

C4.5 classifier is an extension of ID3 algorithm which is used to generate decision trees. C4.5 is a classifier that builds decision trees from training data using information entropy.

The training data set is of classified samples.

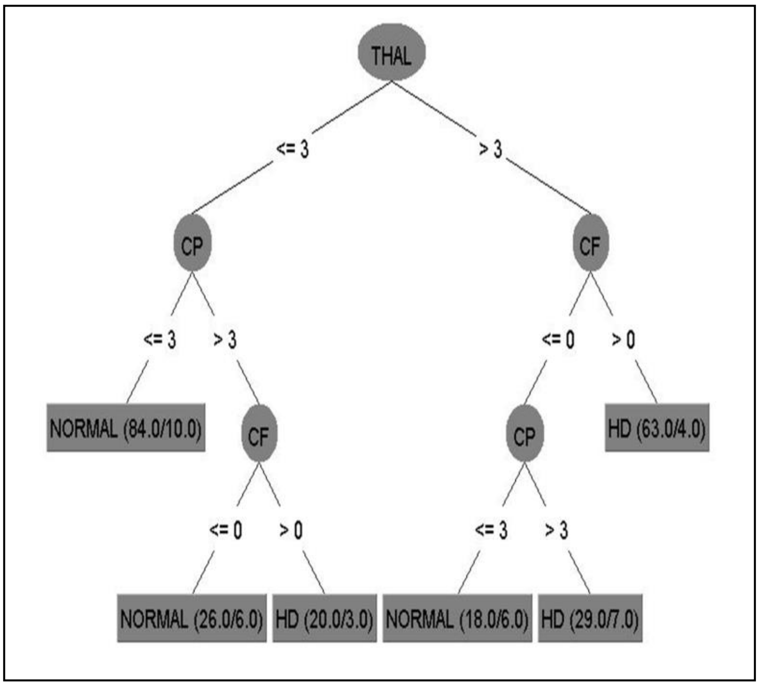


Fig. C4.5 Decision tree

Each sample consists of a dimensional vector, where the attribute values represents the feature of the sample, as well as the class in which it falls.

At each node of the tree, C4.5 chooses the attribute of the data that most selectively splits its set of samples into subsets in one class or the other. The splitting criteria is normalized information gain. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then continues its operations on the smaller sublists.

Naive -Bayes classifier:

The Naïve-Bayes classifier works as follows:

(1) As usual, for an n-dimensional attribute vector, $X = (a_1,a_2,...a_m)$ depicting it from m attributes, respectively $a_1,a_2,.....a_m$.

(2) Suppose that there are n classes: $C=(y_1,y_2,....,y_n)$. Given a tuple, X, the naïve Bayes algorithm predicts that tuple X belongs to class y_i, if and only if $P(y_i \mid X) > P(j \mid X)$ $(1 \leq j \leq n, j \neq i)$. This is called the maximum a posteriori hypothesis.

(3) Thus, using Bayes' theorem, the conditional probability can be decomposed as

$$P\ (yi \mid X) = \frac{P(X\ |yi)P(yi)}{P(X)} = \frac{P(X\ |yi)P(yi)}{\sum_{i=1}^{n} P(X\ |yi)\ P(yi)}$$

(4) This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple. This means that

$$P(X \mid yi) = P(a1\ |yi)\ P(a2 \mid yi) \dots .P(am \mid yi) \dots P(am \mid yi) = \prod P(aj \mid yi)$$

(5) Therefore, formula becomes

$$P(yi \mid X) == \frac{P(X\ |yi)P(yi)}{\sum_{i=1}^{n} P(X\ |yi)\ P(yi)} = \frac{P(yi)\prod P(aj \mid yi)}{\sum_{i=1}^{n} P(yi)\prod P(aj \mid yi)}$$

4. METHODOLOGY

This system uses CRISP-DM i.e. Cross Industry Standard Process for Data Mining for building mining models. There are 6 phases in this method: business understanding, data understanding, data preparation, modelling, evaluation and deployment.

Business understanding looks for focusing on initial requirements, objectives to be fulfilled, taking all the gathered data and transforming it into a problem definition from a business perspective and designing a prefatory plan for achieving necessities stated above.

Data understanding uses the initially gathered data for understanding it and identifying its quality, gain prior acumen and recognize subsets to form a theory.

Data preparation as the name suggests prepares the final dataset that will be sent to modelling tools. This includes all the attributes, tables, records and data cleaning.

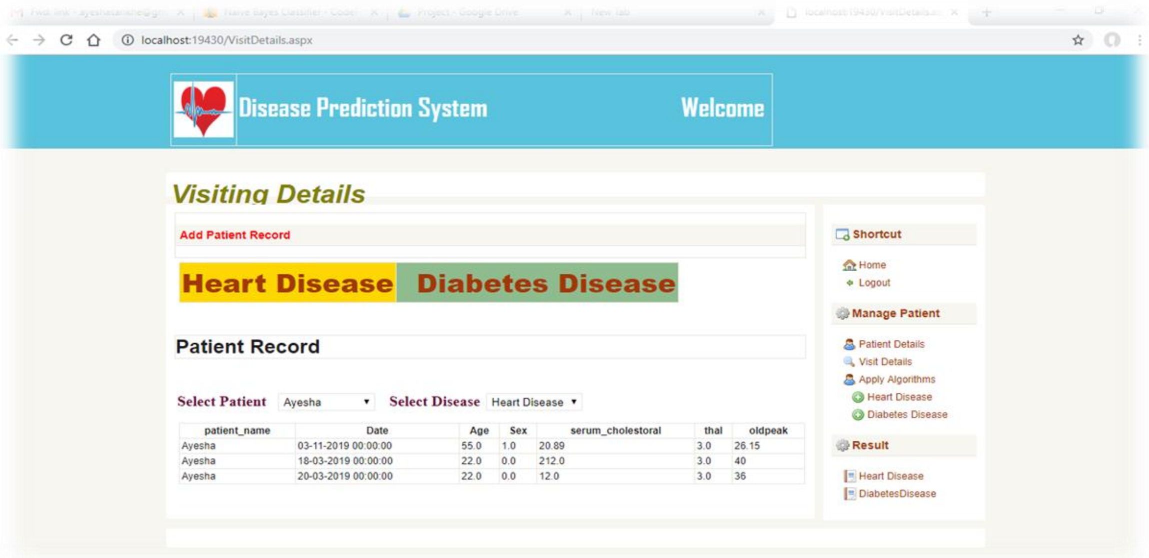
Modelling phase is used for optimization of parameters by using various techniques,

Evaluation phase keeps a tab whether all the objectives stated initially are fulfilled or not.

Deployment phase states all the models that will be required to use data mining models.

Data Mining Extension (DMX), SQL style query language for data mining accesses and builds the contents of any given model. Further tabular and graphical representations are used to get a fuller view of results.

5. EXPERIMENTAL RESULTS



Visiting Details

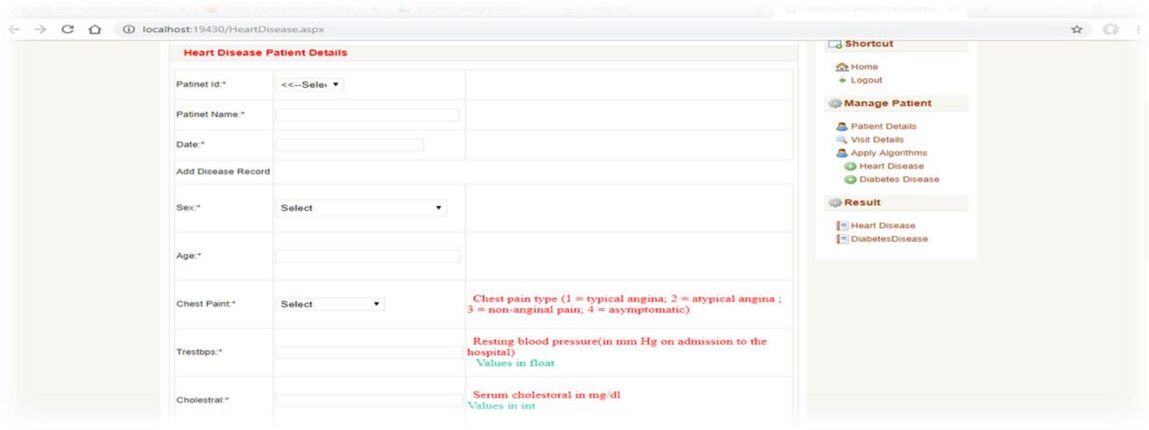
Add Patient Record

Heart Disease Diabetes Disease

Patient Record

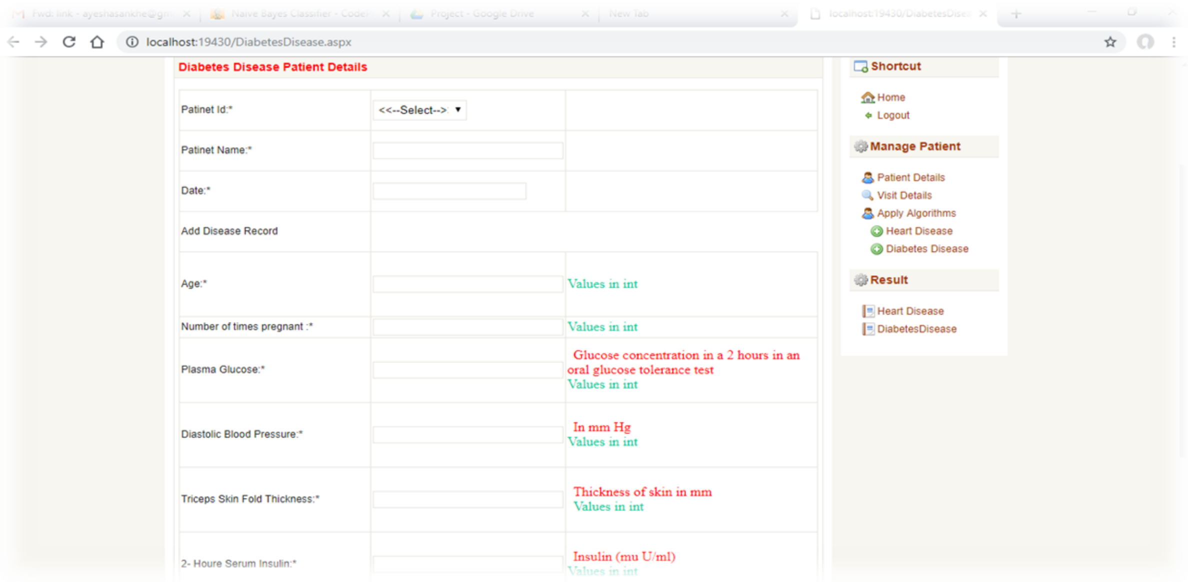
Select Patient: Ayesha ▾ Select Disease: Heart Disease ▾

patient_name	Date	Age	Sex	serum_cholestoral	thal	oldpeak
Ayesha	03-11-2019 00:00:00	55.0	1.0	20.89	3.0	26.15
Ayesha	18-03-2019 00:00:00	22.0	0.0	212.0	3.0	40
Ayesha	20-03-2019 00:00:00	22.0	0.0	12.0	3.0	36



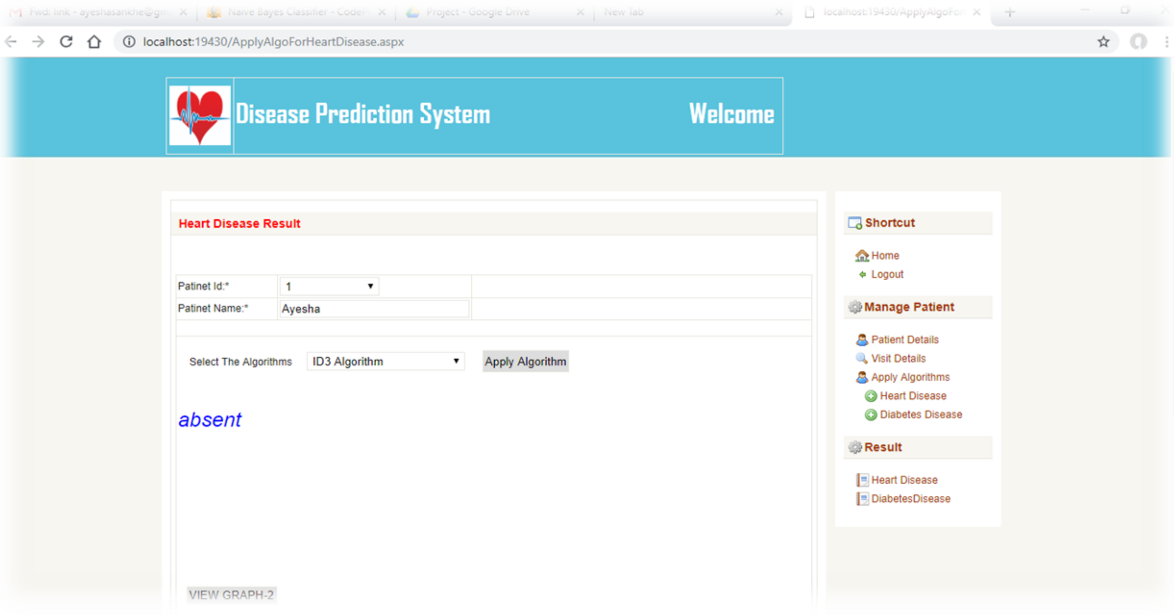
Heart Disease Patient Details

Patinet Id:*	<<--Sele ▾	
Patinet Name:*		
Date:*		
Add Disease Record		
Sex:*	Select ▾	
Age:*		
Chest Paint:*	Select ▾	Chest pain type (1 = typical angina; 2 = atypical angina ; 3 = non-anginal pain; 4 = asymptomatic)
Trestbps:*		Resting blood pressure(in mm Hg on admission to the hospital) Values in float
Cholestral:*		Serum cholestoral in mg/dl Values in int



Diabetes Disease Patient Details

Patinet Id:*	<<--Select--> ▾	
Patinet Name:*		
Date:*		
Add Disease Record		
Age:*		Values in int
Number of times pregnant :*		Values in int
Plasma Glucose:*		Glucose concentration in a 2 hours in an oral glucose tolerance test Values in int
Diastolic Blood Pressure:*		In mm Hg Values in int
Triceps Skin Fold Thickness:*		Thickness of skin in mm Values in int
2- Houre Serum Insulin:*		Insulin (mu U/ml) Values in int

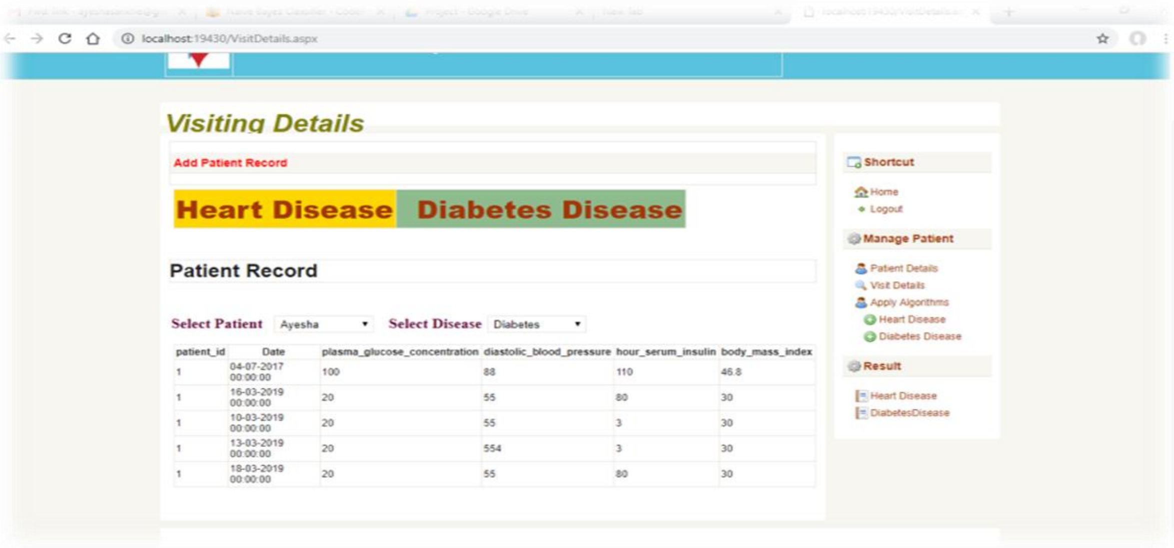


Heart Disease Result

Patinet Id:*	1 ▾
Patinet Name:*	Ayesha

Select The Algorithms ID3 Algorithm ▾

absent



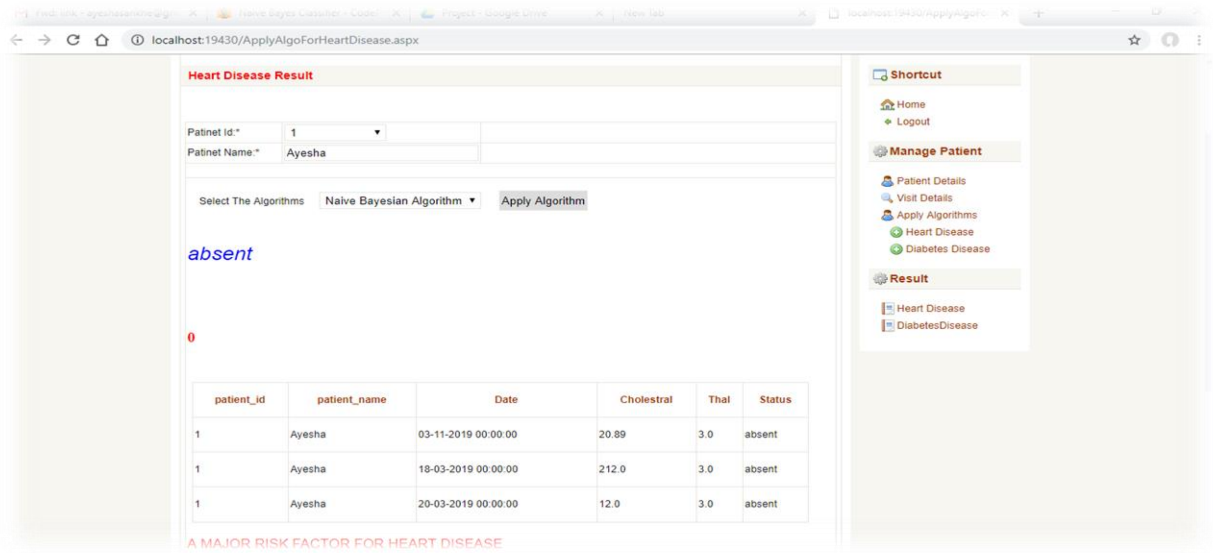
Visiting Details

Add Patient Record

Patient Record

Select Patient Ayesha ▾ Select Disease Diabetes ▾

patient_id	Date	plasma_glucose_concentration	diastolic_blood_pressure	hour_serum_insulin	body_mass_index
1	04-07-2017 00:00:00	100	88	110	46.8
1	16-03-2019 00:00:00	20	55	80	30
1	10-03-2019 00:00:00	20	55	3	30
1	13-03-2019 00:00:00	20	554	3	30
1	18-03-2019 00:00:00	20	55	80	30


localhost:19430/ApplyAlgoForHeartDisease.aspx

Heart Disease Result

Patinet Id:*	1
Patinet Name:*	Ayesha

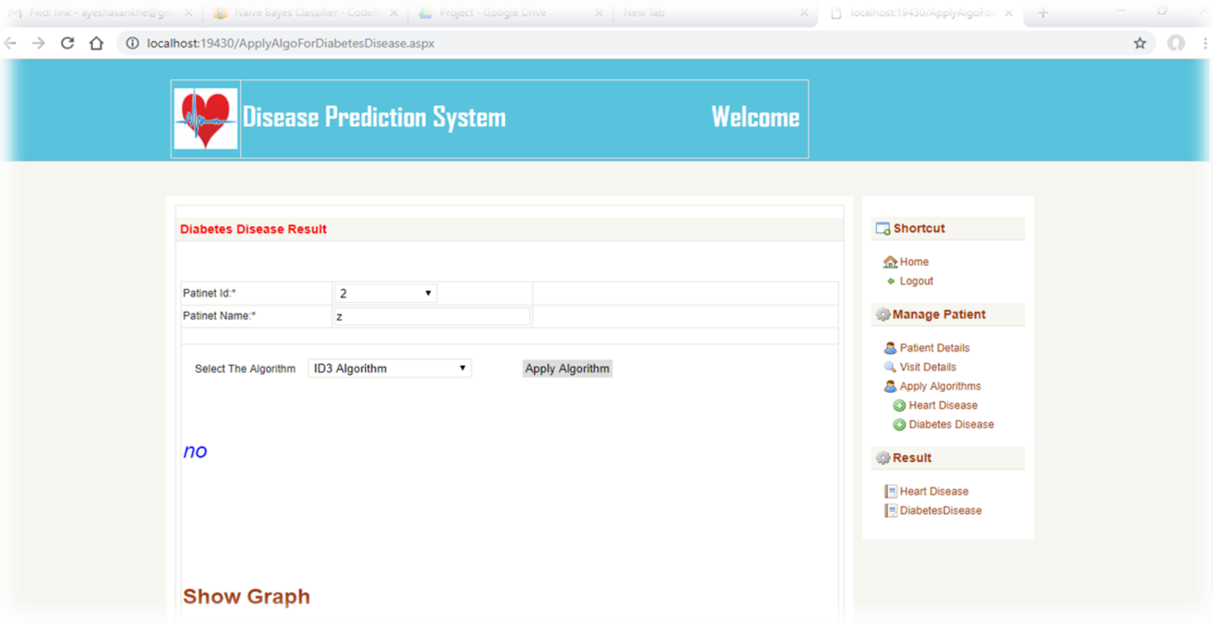
Select The Algorithms: Naive Bayesian Algorithm

absent

0

patient_id	patient_name	Date	Cholestral	Thal	Status
1	Ayesha	03-11-2019 00:00:00	20.89	3.0	absent
1	Ayesha	18-03-2019 00:00:00	212.0	3.0	absent
1	Ayesha	20-03-2019 00:00:00	12.0	3.0	absent

A MAJOR RISK FACTOR FOR HEART DISEASE


localhost:19430/ApplyAlgoForDiabetesDisease.aspx

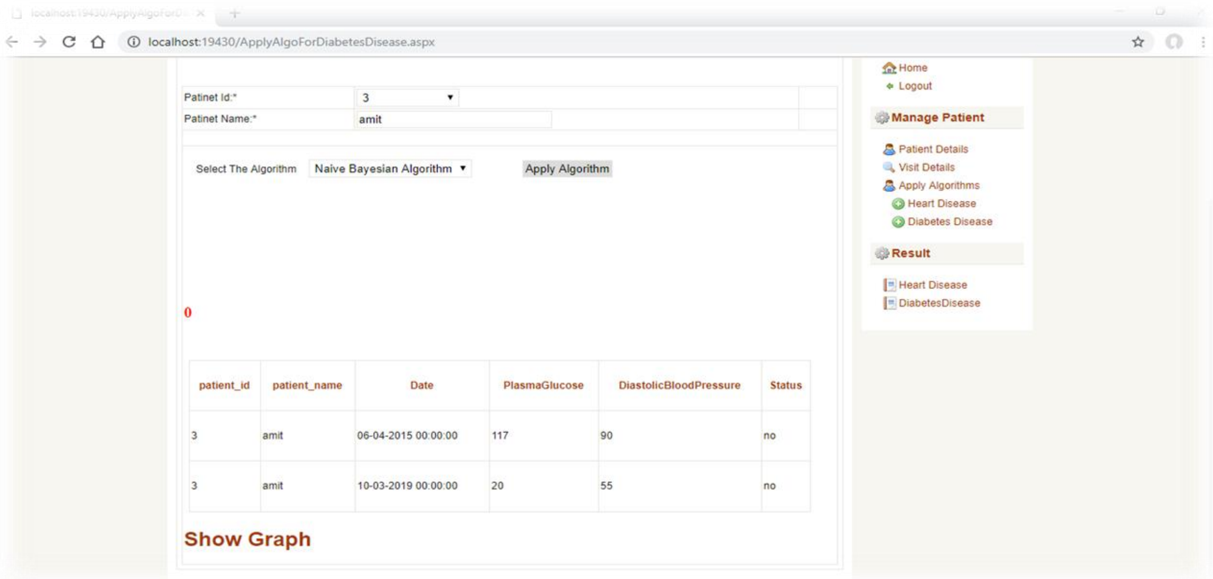
Disease Prediction System Welcome

Diabetes Disease Result

Patinet Id:*	2
Patinet Name:*	z

Select The Algorithm: ID3 Algorithm

no


localhost:19430/ApplyAlgoForDiabetesDisease.aspx

Patinet Id:*	3
Patinet Name:*	amit

Select The Algorithm: Naive Bayesian Algorithm

0

patient_id	patient_name	Date	PlasmaGlucose	DiastolicBloodPressure	Status
3	amit	06-04-2015 00:00:00	117	90	no
3	amit	10-03-2019 00:00:00	20	55	no

6. CONCLUSION

This work examines a real-world problem faced by medical professionals in making apt decisions for patients, based on current patient data and best practices encoded in rule base for scenarios where the data might be missing, where patients could omit or misinterpret their profile.

The project is based on using C4.5 and Naïve Bayes Classifier which processes the data and complexity of running data. C4.5 has a better performance in performance generated rules and accuracy. C4.5 is thus better than ID3 since it is better in induction and generalization of rules. Decisions are then stored in decision support repository. This is currently based on a narrow set of diseases and can be further expanded for medical knowledge.

REFERENCES

[1] Decision Support System For Medical Diagnosis Using Data Mining Ijcsi International Journal Of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011.

[2] Using Data Mining Techniques For Diagnosis And Prognosis Of Cancer Disease, International Journal Of Computer Science, Engineering And Information Technology (Ijcseit), Vol.2, No.2, April 2012

[3] Performance Analysis Of Classification Data Mining Techniques Over Heart Disease Data Base International Journal Of Engineering Science Advanced Technology, Volume-2, Issue-3, 470478.

[4] M.M.Abbasi, S. Kashiyarndi, "Clinical Decision Support Systems: A discussion on different methodologies used in health care International Journal of computer Science and Information security, Vol 8,No 4, 2010.

[5] Abidi and S. Hussain, "Medical knowledge morphing via a semantic web framework CBMS07. Twentieth IEEE International Symposium, 2007.

[6] https://www.ijedr.org/papers/IJEDRCP1403038.pdf