GDPS - General Disease Prediction System

Shratik J. Mishra [1], Albar M. Vasi [2], Vinay S. Menon[3], Prof. K. Jayamalini[4]

[1,2,3,4] *Department of Computer Engineering, Shree L.R. Tiwari college of Engineering, Maharashtra, India*

---***---

Abstract - *The successful application of data mining in highly visible fields like e-business, commerce and trade has led to its application in other industries. The medical environment is still information rich but knowledge weak. There is a wealth of data possible within the medical systems. However, there is a lack of powerful analysis tools to identify hidden relationships and trends in data. Disease is a term that assigns to a large number of heath care conditions related to the body. These medical conditions describe the unexpected health conditions that directly control all the body parts. Medical data mining techniques like association rule mining, classification, clustering is implemented to analyze the different kinds of general body based problems. Classification is an important problem in data mining. A number of popular classifiers construct decision trees to generate class models. The data classification is based on ID3 Decision Tree algorithm which result in accuracy, the data is estimated using entropy based cross validations and partition techniques and the results are compared.*

***Key Words*: Disease, prediction, machine-learning, data mining, ID3.**

1. INTRODUCTION

It is estimated that more than 70% of people in India are prone to general body diseases like viral, flu, cough, cold .etc, in every 2 months. Because many people don't realize that the general body diseases could be symptoms to something more harmful, 25 % of the population succumbs to death because of ignoring the early general body symptoms. This could be a dangerous situation for the population and can be are alarming. Hence identifying or predicting the disease at the earliest is very important to avoid any unwanted casualties. The currently available systems are the systems that are either dedicated to a particular disease or are in research phase for algorithms when it comes to generalized disease.

The purpose of this system is to provide prediction for the general and more commonly occurring disease that when unchecked can turn into fatal disease. The system applies data mining techniques and ID3 decision tree algorithms. This system will predict the most possible disease based on the given symptoms and precautionary measures required to avoid the aggression of disease, it will also help the doctors analyse the pattern of presence of diseases in the society. In this project, the disease prediction system will carry out data mining in its preliminary stages, the system will be trained using machine learning and data mining.

The paper is divided into five sections. The first section gives a brief introduction of about the system. The second section is about data mining and the study of related existing systems. The third section details out the implementation of the system. The fourth section provides the results obtained using mining algorithms. Finally the conclusion gives the summary and future scope about the system.

2. LITERATURE REVIEW

Here we will elaborate the aspects like the literature survey of the project and what all projects are existing and been actually used in the market which the makers of this project took the inspiration from and thus decided to go ahead with the project covering with the problem statement.

2.1 Existing Systems

The authors of this project, Narander Kumar and Sabita Khatri [1], have researched and made comparisons of different algorithms such as k-NN, Naïve Bayes, Random Forest, J48, using performance measures like ROC, kappa statistics, RMSE and MAE in WEKA tools, and also compared the classifiers on various accuracy measures. The conclusion reached of this research was that Random Forest has better accuracy for chronic kidney dataset that was used.

In this project the authors, Monika Gandhi and Dr. Shailendra Singh [2], have analyzed different data mining algorithms like Naïve Bayes, Neural network and decision tree algorithms for their accuracy on prediction of Heart Disease.

The authors Marija Sultana, Afrin Haider and Md.Shorif Uddin [3], have analyzed algorithms such as K-star, J48, SMO, Bayes Net and Multilayer Perceptron Network using WEKA tools for heart disease prediction dataset. The performance of these datamining techniques in acquired by combination of results of measures such as predictive accuracy, ROC curve and AUC value. The result obtained is the SMO and Bayes network show more optimum result than their other mentioned counterparts.

In this project, the authors Girija D.K, Dr. M.S. Shashidhara and M.Giri [4], make use of Neural networks to make predictions regarding presence of uterine fibroid disease. The experimental results show an accuracy of 98% using the Multilayer perceptron neural network and data mining.

This project focuses on the most common form of cancer present in women i.e. Breast Cancer and its recurrence. The authors Uma Ojha and Dr. Savita Goel, in this project have researched on many data mining algorithms in both classification and clustering types. They reached an observed conclusion that, C5.0 and Fuzzy C-means provide highest and lowest accuracy amongst all the used algorithms, respectively. While C5.0 is highest with 81% accuracy on the given data, Fuzzy C-means is lowest with 37% accuracy. The other generalized conclusion reached is that Classification algorithms provide better accuracy over Clustering algorithms [5].

3. SYSTEM ARCHITECTURE.

The general disease prediction system predicts chance of presence of a disease present in a patient on the basis of their symptoms. It will also recommend necessary precautionary measures required to treat the predicted disease. The system will initially be fed data from different sources i.e. patients, the data will then be pre-processed before further process is carried out, this is done so as to get clean data from the raw initial data, as the raw data would be noisy, or flawed. This data will be processed using Data mining algorithms, the system, will be trained so as to predict the disease based on the input data given by the user.

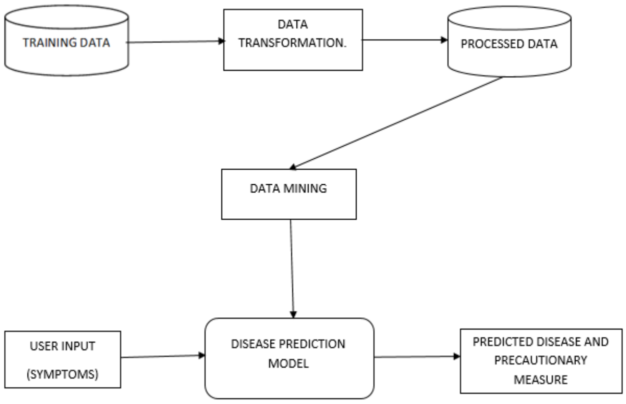


Fig -1: Block Diagram for General disease prediction system.

The system is implemented into two parts, admin part and the user part. The duty of the admin is training the system for creation of the disease prediction model. The user uses the services provided by the model after logging in as the user, entering the symptoms into the model, which in turn returns the predicted results and necessary precautionary measures.

4. IMPLEMENTATION

The system is divided into two parts, an admin part and the user part.

4.1 Admin

After logging in the system as verified admin, the admin will carry out following duties.

a. Data Preparation.

b. Data Transformation.

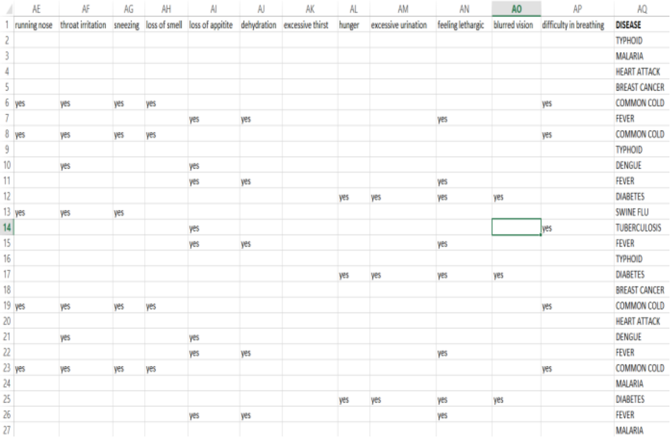
c. Feature Extraction.

d. Implementation of ID3 algorithm.

e. Model.

Step 1: Data Preparation

The data set obtained in this study was obtained from a local hospital based in Mumbai. Initially the size of data was 120.



	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ
1	running nose	throat irritation	sneezing	loss of smell	loss of appitite	dehydration	excessive thirst	hunger	excessive urination	feeling lethargic	blurred vision	difficulty in breathing	**DISEASE**
2													TYPHOID
3													MALARIA
4													HEART ATTACK
5													BREAST CANCER
6	yes	yes	yes	yes								yes	COMMON COLD
7					yes	yes				yes			FEVER
8	yes	yes	yes	yes								yes	COMMON COLD
9													TYPHOID
10		yes			yes								DENGUE
11					yes	yes				yes			FEVER
12								yes	yes	yes	yes		DIABETES
13	yes	yes	yes										SWINE FLU
14					yes							yes	TUBERCULOSIS
15					yes	yes				yes			FEVER
16													TYPHOID
17								yes	yes	yes	yes		DIABETES
18													BREAST CANCER
19	yes	yes	yes	yes								yes	COMMON COLD
20													HEART ATTACK
21		yes			yes								DENGUE
22					yes	yes				yes			FEVER
23	yes	yes	yes	yes								yes	COMMON COLD
24													MALARIA
25								yes	yes	yes	yes		DIABETES
26					yes	yes				yes			FEVER
27													MALARIA

Fig -2: Data taken

Step 2: Data Transformation

In this step, the dataset is explored and necessary data is selected and the dataset is converted into machine understandable form.

Step 3: Feature Extraction

Feature extraction is the process to reduce the size of data so as to only take informative, non-redundant and relevant data, so as to facilitate subsequent learning and generalization step to acquire better human interpretation

Step 4: Implementation of ID3 algorithm

The ID3 algorithm is used in this project, for training of the disease prediction model. When the algorithm is applied, it generates a rule set based on the observed pattern of data. On the basis of this rule set, the system training is done and the model is created.

ID3 algorithm:

The project makes use of ID3 algorithm. ID3 stands for Iterative Dichotomiser 3. The algorithm was developed by Ross Quinlan and is used to generate a decision tree from a given dataset. ID3 works mainly on three things, firstly the entropy of each attribute, second information gain and third, entropy of whole dataset, using these three, it picks a root node. The condition for selection of root node is that, the attribute with lower value of entropy (or higher value of information gain.) becomes the root node. This carries on until the last element of data that can provide some substantial information is not used.

- Calculate the entropy of every attribute using the data set S.
- Split the set S into subsets using the attribute for which the resulting entropy (after splitting) is minimum (or, equivalently, information gain is maximum)
- Make a decision tree node containing that attribute
- Recurse on subsets using remaining attributes.

Step5: Model Creation

Once the system training is done, our general body disease prediction model will be ready for using,

4.2 USER

On logging in the system as user, the user can carry out following functionalities of the system.

a. Entering Symptoms

b. Disease Prediction

c. Precautions

Step 1: Entering Symptoms

User once logged in can select the symptoms presented by them, available in the drop-down box.

Step 2: Disease prediction

The predictive model predicts the disease a person might have based on the user entered symptoms.

Step 3: Precautions

The system also gives required precautionary measures to overcome a disease.

5. RESULTS

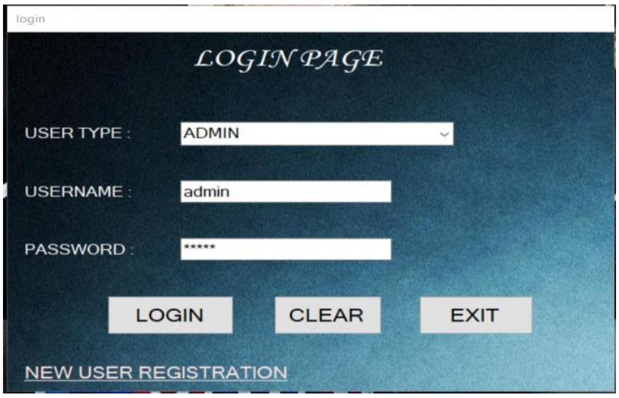
The General body disease prediction system applies data mining techniques using ID3 algorithm. Decision trees are considered easily understood models because a reasoning process can be given for each conclusion. Knowledge models under this paradigm can be directly transformed into a set of IF-THEN rules that are one of the most popular forms of knowledge representation.

5.1 Admin

The results for the admin of the GDPS are as follows.

Step 1: Login

Admin can login the system selecting the user type and entering the required details.



login

LOGIN PAGE

USER TYPE : ADMIN

USERNAME : admin

PASSWORD : *****

Fig -3: Login Page for admin in GDPS

Step 2: System Training

The admin should train the system by uploading the data set into the system.

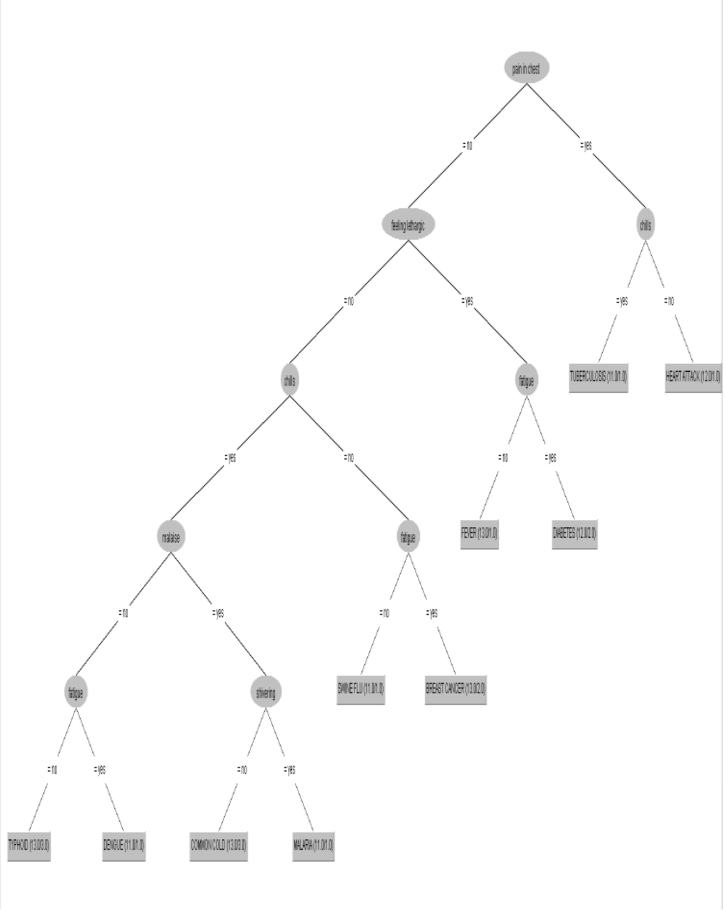
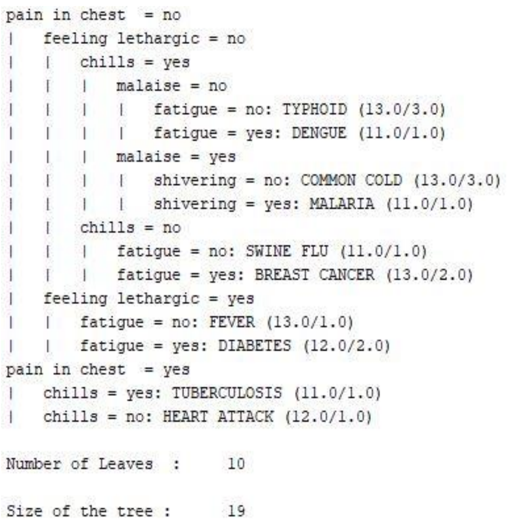


Fig -4: Tree structure of GDPS



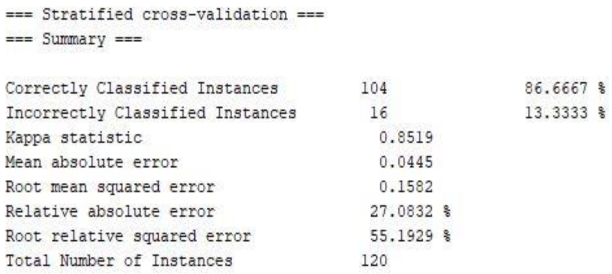
```
pain in chest  = no
|   feeling lethargic = no
|   |   chills = yes
|   |   |   malaise = no
|   |   |   |   fatigue = no: TYPHOID (13.0/3.0)
|   |   |   |   fatigue = yes: DENGUE (11.0/1.0)
|   |   |   malaise = yes
|   |   |   |   shivering = no: COMMON COLD (13.0/3.0)
|   |   |   |   shivering = yes: MALARIA (11.0/1.0)
|   |   chills = no
|   |   |   fatigue = no: SWINE FLU (11.0/1.0)
|   |   |   fatigue = yes: BREAST CANCER (13.0/2.0)
|   feeling lethargic = yes
|   |   fatigue = no: FEVER (13.0/1.0)
|   |   fatigue = yes: DIABETES (12.0/2.0)
pain in chest  = yes
|   chills = yes: TUBERCULOSIS (11.0/1.0)
|   chills = no: HEART ATTACK (12.0/1.0)

Number of Leaves  :     10

Size of the tree :      19
```

Fig -5: Pruned tree structure of GDPS.

Experiments were carried out in order to evaluate the performance and usefulness of different classification algorithms for predicting disease present in patient. The results of the experiments are shown below:

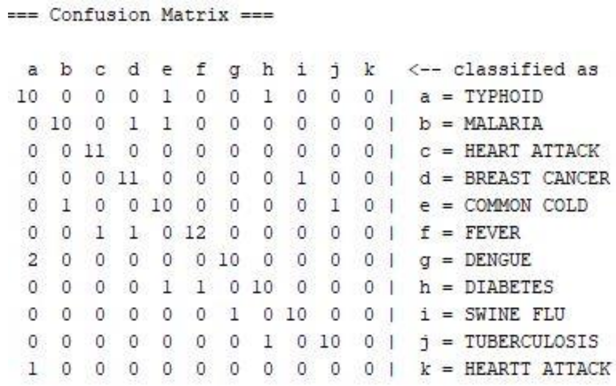


```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         104               86.6667 %
Incorrectly Classified Instances        16               13.3333 %
Kappa statistic                          0.8519
Mean absolute error                      0.0445
Root mean squared error                  0.1582
Relative absolute error                 27.0832 %
Root relative squared error             55.1929 %
Total Number of Instances              120
```

Fig -6: Stratified cross validation for GDPS.

The performance of the learning techniques is highly dependent on the nature of the training data. Confusion matrices are very useful for evaluating classifiers. The columns represent the predictions, and the rows represent the actual class.



```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k   <-- classified as
 10  0  0  0  1  0  0  1  0  0  0 |  a = TYPHOID
  0 10  0  1  1  0  0  0  0  0  0 |  b = MALARIA
  0  0 11  0  0  0  0  0  0  0  0 |  c = HEART ATTACK
  0  0  0 11  0  0  0  0  1  0  0 |  d = BREAST CANCER
  0  1  0  0 10  0  0  0  0  1  0 |  e = COMMON COLD
  0  0  1  1  0 12  0  0  0  0  0 |  f = FEVER
  2  0  0  0  0  0 10  0  0  0  0 |  g = DENGUE
  0  0  0  0  1  1  0 10  0  0  0 |  h = DIABETES
  0  0  0  0  0  0  1  0 10  0  0 |  i = SWINE FLU
  0  0  0  0  0  0  0  1  0 10  0 |  j = TUBERCULOSIS
  1  0  0  0  0  0  0  0  0  0  0 |  k = HEARTT ATTACK
```

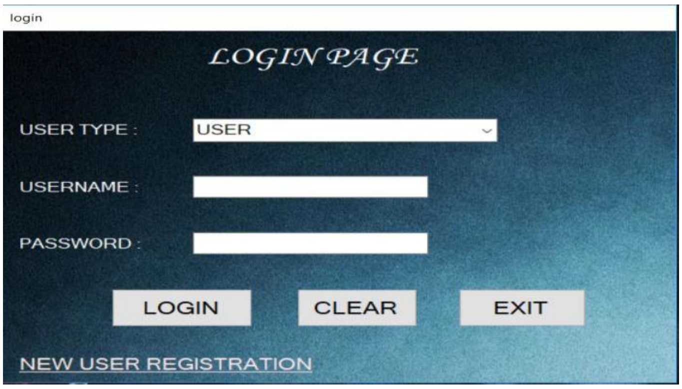
Fig -7: Confusion Matrix of GDPS.

5.2 User

The results for the user in GDPS are as follows.

Step 1: User login

Pre-registered user should login the system to have access to the services.

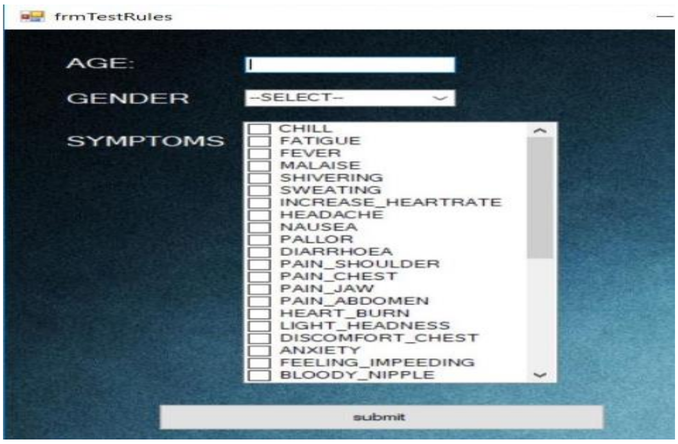


USER TYPE : USER

Fig -8: User Login page for GDPS.

Step 2: Enter Symptoms

User will have to select the symptoms here.



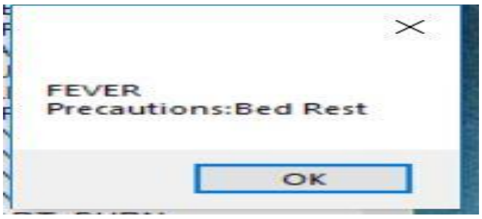
GENDER --SELECT--

□ CHILL
□ FATIGUE
□ FEVER
□ MALAISE
□ SHIVERING
□ SWEATING
□ INCREASE_HEARTRATE
□ HEADACHE
□ NAUSEA
□ PALLOR
□ DIARRHOEA
□ PAIN_SHOULDER
□ PAIN_CHEST
□ PAIN_JAW
□ PAIN_ABDOMEN
□ HEART_BURN
□ LIGHT_HEADNESS
□ DISCOMFORT_CHEST
□ ANXIETY
□ FEELING_IMPEEDING
□ BLOODY_NIPPLE

Fig -9: Symptom selection form for user in GDPS

Step 3: Prediction and precaution

The result calculated by the model based on the rule set will be displayed here.



FEVER
Precautions:Bed Rest

Fig -10: Prediction result for GDPS.

6. CONCLUSIONS:

The system has been implemented with the accuracy of 86.67% on the dataset of 120 patient data. The current system covers only the general diseases or the more commonly occurring disease, the plan is to include disease of higher fatality, like various cancers in future, so that early prediction and treatment could be done, and the fatality rate of deadly diseases like cancer decreases, with the economic benefit in long sight as well.

REFERENCES

[1] Implementing WEKA for medical data classification and early disease prediction. “3rd IEEE International Conference on "Computational Intelligence and Communication Technology" (IEEE-CICT 2017)”.

[2] Predictions in Heart Disease Using Techniques of Data Mining, “2015 1st International Conference on Futuristic trend in Computational Analysis and Knowledge Management (ABLAZE-2015)”.

[3] Analysis of Data Mining Techniques for Heart Disease Prediction, “Marjia Sultana, Afrin Haider and Mohammad Shorif Uddin”.

[4] Data mining approach for prediction of fibroid Disease using Neural Networks,” Dr. M.S. Shashidhara, M. Giri, Girija D.K.”

[5] Study on prediction of Breast cancer recurrence using Data mining techniques, “Uma Ojha, Dr. Savita Goel.”

[6] Classification model of Prediction for placement of students, “Saurabh Pal.”