HEALTH PREDICTION ANALYSIS USING DATA MINING

Kritika Ashok Rane
Department of Information Technology
Konkan Gyanpeeth college of Engineering,
Karjat,India
kritika2rane@gmail.com

Ashish Ravindra Salve
Department of Information Technology
Konkan Gyanpeeth college of Engineering,
Karjat, India
a.r.salve@live.in

Ashwini Dhananjay Sawant
Department of Information Technology
Konkan Gyanpeeth college of Engineering,
Karjat, India
sawantashwini02@gmail.com

Data mining techniques are used for a variety of applications. In healthcare industry, datamining plays an important role in predicting diseases. For detecting a disease number of tests should be required from the patient. But using data mining technique the number of tests can be reduced. This reduced test plays an important role in time and performance. This report analyses data mining techniques which can be used for predicting different types of diseases. This report reviewed the research papers which mainly concentrate on predicting various disease

INTRODUCTION

There was a time when data were not readily available. As data become more abundant ,however limitations in computational capabilities prevented the practical application of mathematical models. Consequently, data mining tools are now being used for clinical data. The application of Data Mining in healthcare has a lot of positive and also life-saving outcomes. Data Mining refers to the vast quantities of information created by the digitization of everything, that gets consolidated and analyzed by specific technologies. Applied to healthcare, it will use specific health data of a population (or of a particular individual) and potentially help to prevent epidemics, cure disease, cut down costs, etc. Data mining is a process of selecting, exploring and modelling large amounts of data. This process has become an increasingly pervasive activity in all areas medical science research. Data mining has resulted in tha discovery of useful hidden patterns from massive databases. By using data mining techniques finally physicians need to know how quickly identify and diagnose potential cases

Data Mining:-

Data mining is a process of selecting, exploring and modelling large amounts of data. Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. The major steps involved in a data mining process are: Extract, transform and load data into a data warehouse. Store and manage data in a multidimensional databases. Provide data access to business analysts using application software. Present analysed data in easily understandable forms, such as graphs.

What Is The Use Of Data Mining In Healthcare?

The application of Data Mining in healthcare has a lot of positive and also life-saving outcomes. Data Mining refers to the vast quantities of information created by the digitization of everything, that gets consolidated and analyzed by specific technologies. Applied to healthcare, it will use specific health data of a population (or of a particular individual) and potentially help to prevent epidemics, cure disease, cut down costs, etc

Objectives:-

1. Predictive Analytics And Preventive Measures
2. The Ultimate HER
3. Disease Modelling and Mapping
4. Reduce Fraud And Enhance Security
5. Personalized Medicine

LITERATURE SURVEY

In the paper *"Data mining application in health care sector" by M.Sumanth* he to aimed to compare the different data mining application in the healthcare sector for extracting useful information. The prediction of disease data mining applications but it is challenging It need human efforts and improve accuracy. He also mention the application of data mining in healthcare sector such as Treatment effectiveness, Healthcare management,Customer relationship management, Medical device industry, Pharmaceutical industry, Hospital management, System biology.

In "Hybrid Approach for Heart Disease Detection Using Clustering and ANN" by Tejasweeta Dixit ,Reshmi Gore, Prerana Akade they stated that the Heart disease is dominant caused death in developed countries and main contributors to disease strain in develop countries. Data mining the extraction of hidden predictive from large database is powerful new technology with great probability to help companies focus on the most important in info in their data warehouse. By using various data mining techniques such as clustering algorithem,decision tree,classifiers, we can predict heart diseases

As stated in *"Application of big data in medical science brings in revolution in managing health care of humans" by Dr.Gagandeep Jagdev, Sukhpreet Singh* By enabling research to identify compounds with higher likelihood of success big data can help reduce the cost and the time to market for diagnosing the diseased. They also mentioned the roles played by big data in medical science such as Personalized treatment planning, Assisted diagnosis ,Fraud direction ,Monitor patient vital signs, Digitization of disease.

As In the report *"Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset"* by *Tapas Ranjan Baitharua , Subhendu Kumar Panib* they conducted experiment on Liver Disorder Dataset by using WEKA Tool and they had found the impact of liver disorder on the predictive performance of different classifiers. after they analyzed quantitative data they found that we find that the general concept of improved predictive performance of all above classifiers but Naive Bayes performance is not significant.

IMPLEMENTATION METHODOLOGIES

1. K-nearest neighbors algorithm:-

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition. Choosing the optimal value for K is best done by first inspecting the data. In general, a large K value is more precise as it reduces the overall noise but there is no guarantee. Cross-validation is another way to retrospectively determine a good K value by using an independent dataset to validate the K value. Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN

2. Logistic Regression :-

Logistic regression predicts the probability of an outcome that can only have two values (i.e. a dichotomy). The prediction is based on the use of one or several predictors (numerical and categorical). A linear regression is not appropriate for predicting the value of a binary variable for two reasons:

- A linear regression will predict values outside the acceptable range (e.g. predicting probabilities outside the range 0 to 1)
- Since the dichotomous experiments can only have one of two possible values for each experiment, the residuals will not be normally distributed about the predicted line.

On the other hand, a logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression is similar to a linear regression, but the curve is constructed using the natural logarithm of the "odds" of the target variable, rather than the probability. Moreover, the predictors do not have to be normally distributed or have equal variance in each group.

3. Random forest :-

Random forests provide an improvement over bagged trees by way of a random forest small tweak that decorrelates the trees. As in bagging, we build a number of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors. The split is allowed to use only one of those m predictors. A fresh sample of m predictors is taken at each split, and typically we choose $m \approx \sqrt{p}$—that is, the number of predictors considered at each split is approximately equal to the square root of the total number of predictors (4 out of the 13 for the Heart data).

SYSTEM DESIGN

We had followed the divide and conquer theory, so we divided the overall problem into 3 parts and develop each part or module separately. When all modules are ready, we should integrate all the modules into one system. We briefly described all the modules and the functionality of these modules bellow:

- **User Authentication Module**

This module will provide functionality of user authentication, creation manages the credentials for user to authenticate.

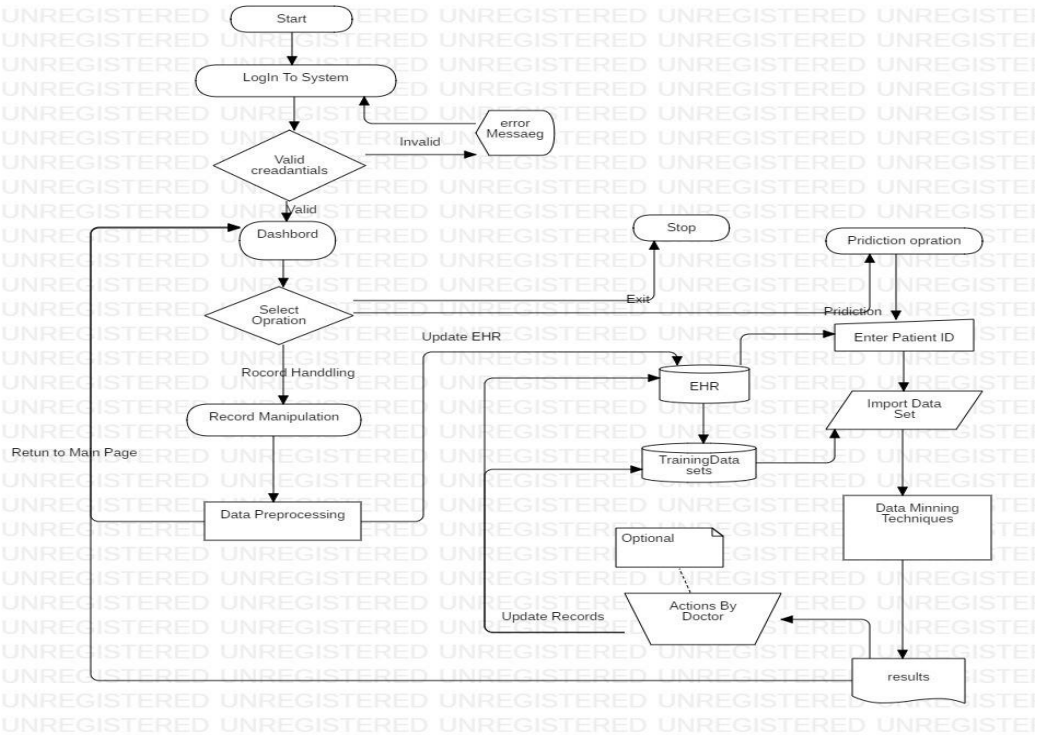
- **Data Manipulation Module**

This module will provide functionality of creation, deletion, and modification of patient data. This module will also perform the data pre-processing since data provided by user may contain some missing field and noisy data which needs to be pre-processed.

- **Prediction And Results Module**

This module will Provide the prediction functionality for the system ,the prediction is done by using various data mining techniques. The prediction results were also handled by this module.

Flow of System



CALCULATION REPORTS FOR VARIOUS ALGORITHM TESTED

For finding the confusion matrix and accuracy of the trained model we used same dataset named as heart_disease in that there were few parameters regarding the heart disease and the final outcome which indicates actual result i.e. whether the patient had a heart disease or not so, by using this data we had trained a model using python & sklearn library.

For calculating confusion matrix, we required some training data and test data for that We split this Dataset into two different datasets, one for the independent Attributex, and one for the Outcome Attribute y (which is the last column). We'll now split the dataset x into two separate setsxTrain and xTest. Similarly, we'll split the dataset y into two sets as well yTrain and yTest. we have split the dataset in a 70–30 ratio, by specifying the test size parameter to 0.3

Then for generating the following Confusion matrix We had used classification_report() functions from sklearn library with parameters Ground truth (correct) target values i.e. yTest and Obtained Prediction value by using respective algorithms. Then we find the accuracy by using accuracy_score() function with same parameters as Confusion matrix

Following are the Output of tests on various algorithms which we had implemented. The output contains confusion matrix and calculated accuracy

```
-------------------logistic regression --------------
confusion matrix
precision    recall  f1-score   support
1       0.81      0.75      0.78        40
2       0.77      0.83      0.80        41
avg / total       0.79      0.79      0.79        81

accuracy of logistic regression is
0.7901234567901234

----------------Random Forest Classifier---------------
confusion matrix
precision    recall  f1-score   support
1       0.77      0.75      0.76        40
2       0.76      0.78      0.77        41
avg / total       0.77      0.77      0.77        81

accuracy of random forest is
0.7654320987654321

------------- K Neighbours Classifier-----------------------
confusion matrix
precision    recall  f1-score   support
1       0.61      0.75      0.67        40
2       0.69      0.54      0.60        41
avg / total       0.65      0.64      0.64        81

accuracy of KNN is
0.6419753086419753
```

LIMITATIONS OF THE SYSTEM

1. Privacy at risk since data is available to anyone who is authenticated or it may happen that the data is unintentionally got visible to an unauthorized user and he may miss-use your data.

2. Requires large training set the system requires large no of data in training set to predict accurately,, more the data the more accurate prediction is

3. Data manipulation is difficult and time consuming since it require large no of data filed or parameters, the system has to be able to manage it.for managing that it require more time and much more complex dataset.

FUTURE SCOPE OF THE PROJECT

In future the system can be used as:-

1. As an mobile based application.

2. In future the system can work Can work with medical electronics to keep monitoring and analysing patients health in real time.

3. In future the system can work As a nation wide health database attached with patients UID(adhar).

4. In future the system can work with voice assistants and AI.So users doesn't need to always keep track of their reports since the AI can automatically does it for user and voice assistance can also provide day to day medical feeds to user

CONCLUSION

The system can get great importance for area of healthcare, and it represents comprehensive process that demands thorough understanding of needs of the healthcare organizations. Knowledge gained with the use of the system can be used to make successful decisions that will improve success of healthcare organization and health of the patients. the system requires appropriate technology and analytical techniques, as well as systems for reporting and tracking which can enable measuring of results.

ACKNOWLEDGMENT

Success is nourished under the combination of perfect guidance, care blessing. Acknowledgement is the best way to convey. We express deep sense of gratitude brightness to the outstanding permutations associated with success. Last few years spend in this estimated institution has molded us into confident and aspiring Engineers. We express our sense of gratitude towards our project guide Prof. J. P. Patil. It is because of his valuable guidance, analytical approach and encouragement that we could learn and work on the project. We will always cherish the great experience to work under their enthusiastic guidance.

REFERENCES

1. M. Sumanth. *Data Mining Applications in Healthcare Sector*.

2. Neha Chikshe, Tejasweeta Dixit, Rashmi Gore Prerana Akade (2016). Hybrid Approach for Heart Disease Detection Using Clustering and ANN. *IJRITCC,JAN 2016 Volume 4 Issue 1*.

3. Dr. Gagandeep Jagdev. Application Of Big Data In Medical Science Brings Revolution In Managing Health Care Of Humans. *IJEEE,JAN 2015 Volume 2 SPl. Issue 1*

4. Gareth James, Daniela Witten ,Trevor Hastie , Robert Tibshirani ,"An Introduction to Statistical Learning"

5. Tapas Ranjan Baitharua , Subhendu Kumar Panib "Analysis of Data Mining Techniques For Healthcare Decision Support System Using Liver Disorder Dataset". *International Conference on Computational Modeling and Security (CMS 2016)*

6. S.K. Shinde & Uddagiri Chandrashekhar "Data Mining and business Intelligence "

Websites:

1. https://ijritcc.com/index.php/ijritcc/article/view/1718
2. https://www.slideshare.net/madallapallisumanth/data-mininginhealthcaresector
3. https://issuu.com/ijeeeapm/docs/id77
4. https : //www.immagic.com/eLibrary/ARCHIV ES/GENERAL/W IKIP EDI/W 1120615B.pdf
5. www.W ikipedia.com
6. https://www.saedsayad.com
7. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html
8. https://www.python.org